Videos in Context for Telecommunication and Spatial Browsing

Fabrizio Pece

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

Department of Computer Science
University College London

January 19, 2015

I, Fabrizio Pece, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

A person who never made a mistake never tried anything new.

Albert Einstein

To my parents.

Abstract

The research presented in this thesis explores the use of videos embedded in panoramic imagery to transmit spatial and temporal information describing remote environments and their dynamics. Virtual environments (VEs) through which users can explore remote locations are rapidly emerging as a popular medium of presence and remote collaboration. However, capturing visual representation of locations to be used in VEs is usually a tedious process that requires either manual modelling of environments or the employment of specific hardware. Capturing environment dynamics is not straightforward either, and it is usually performed through specific tracking hardware. Similarly, browsing large unstructured video-collections with available tools is difficult, as the abundance of spatial and temporal information makes them hard to comprehend. At the same time, on a spectrum between 3D VEs and 2D images, panoramas lie in between, as they offer the same 2D images accessibility while preserving 3D virtual environments surrounding representation. For this reason, panoramas are an attractive basis for videoconferencing and browsing tools as they can relate several videos temporally and spatially.

This research explores methods to acquire, fuse, render and stream data coming from heterogeneous cameras, with the help of panoramic imagery. Three distinct but interrelated questions are addressed. First, the thesis considers how spatially localised video can be used to increase the spatial information transmitted during video mediated communication, and if this improves quality of communication. Second, the research asks whether videos in panoramic context can be used to convey spatial and temporal information of a remote place and the dynamics within, and if this improves users' performance in tasks that require spatio-temporal thinking. Finally, the thesis considers whether there is an impact of display type on reasoning about events within videos in panoramic context. These research questions were investigated over three experiments, covering scenarios common to computer-supported cooperative work and video browsing. To support the investigation, two distinct video+context systems were developed.

The first telecommunication experiment compared our videos in context interface with fully-panoramic video and conventional webcam video conferencing in an object placement scenario. The second experiment investigated the impact of videos in panoramic context on quality of spatio-temporal thinking during localization tasks. To support the experiment, a novel interface to video-collection in panoramic context was developed and compared with common video-browsing tools. The final experimental study investigated the impact of display type on reasoning about events. The study explored three adaptations of our video-collection interface to three display types. The overall conclusion is that videos in panoramic context offer a valid solution to spatio-temporal exploration of remote locations. Our ap-

proach presents a richer visual representation in terms of space and time than standard tools, showing that providing panoramic contexts to video collections makes spatio-temporal tasks easier. To this end, videos in context are suitable alternative to more difficult, and often expensive solutions. These findings are beneficial to many applications, including teleconferencing, virtual tourism and remote assistance.

Acknowledgements

The research presented in this thesis was supported by the BEAMING project[1], which is funded by the European Commission under the FP7 ICT Work Programme.

Firstly, I would like to thank Prof. Jan Kautz, my doctoral supervisor and academic mentor, whose support, knowledge and enthusiasm made my studies and research a lot easier and more enjoyable. I could not have asked for a better supervisor than Jan, who provided me with great guidance through my studies at UCL, while leaving room for my personal ideas. I would also like to thank Dr. Tim Weyrich, my secondary supervisor, and Prof. Anthony Steed, who provided valuable points of reflection and helped developing my ideas and research skills.

Thanks also to my good friend and colleague Dr. Will Steptoe, who contribute to shape my research with his support and who was always ready to help and share a laugh. A special thanks goes also to Dr. James Tompkin, an invaluable friend and colleague whose help and support have helped me in this journey. I am also grateful to all other colleagues and friends from UCL, Disney Research Zürich and ETH Zürich (especially those sharing my office), who provided a highly inspiring environment.

Finally, and most importantly, I want to thank my family and friends for their love and support. To my parents, whose encouragement, love and exemplary helped me getting this far. They are the pillars of my education, and I could not be more grateful to them. To Jana, whose presence next to me makes the happy moments even happier, and the hard ones easier to overcome. Thank you for patiently listening to my (not always so exciting) “research ideas”, and for making my life so much more interesting. Thanks also to Marco, Gerardo, Alessandro and David, my long-standing friends with whom I have shared countless happy and memorable moments.

[1] www.beaming-eu.org

Contents

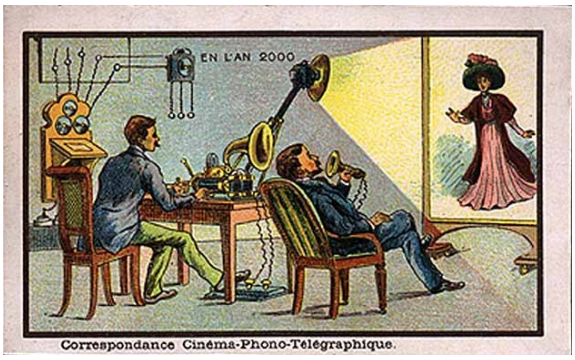
List of Figures

List of Tables

Chapter 1

Introduction

(a) *"Concerts and Opera at Home" trade card from the "One Hundred Years Hence" series. This card shows people with individual receivers listening at home to a live concert, while a funnel-shaped device transfers a visual image to the wall. This predicts the future invention of simultaneous transmission of both sound and picture.*



(b) *"Correspondance Cinema-Phono-Telegraphique" trade card from the "In the Year 2000" (*En L'an 2000*) series. This card shows an invention of the twentieth century with which people are able to communicate through video and sound. The "invention" closely resembles modern video-mediated communication system.*

Figure 1.1: *Two trade cards depict future technological developments as imagined by the mid nineteenth century (*Left*) and beginning of twentieth century (*Right*).*

Ever since the introduction of telephone, mankind has always been fascinated by the opportunity of transmitting video and sound remotely, to allow video-mediated communication (VMC) and "virtual" exploration of remote locations. Generations of researchers have investigated the possibility of transmitting audio-video recordings, for both recreational (Figure 1.1(a)) or communicational (Figure 1.1(b)) purposes. When finally on March 10th 1876 Alexander G. Bell established the first telephone communication with Thomas A. Watson, his first words were "*Mr. Watson, come here, I want to see you*".

Audio telecommunication between geographically-remote people has become a ubiquitous part of life throughout the world. However, only with the introduction in 1964 of the AT&T *Picturephone* [Mol69], Bells wish to see Watson, as well as speak to him, may have been granted. Since then, computing performance has experienced rapid advancement, and so did the possibility to capture and transmit videos in real-time. During the same years, the increase in computing performance allowed researchers to start investigating the field of real-time computer graphics (CG), and when in 1965 Ivan Sutherland

presented *The Ultimate Display* [Sut65], an essay on emergent CG technologies, he envisioned future scenarios which are today a reality:

Don't think of that thing as a screen, think of it as a window, a window through which one looks into a virtual world. The challenge to computer graphics is to make that virtual world look real, sound real, move and respond to interaction in real time... and even feel real!

Sutherland is a pioneer of virtual reality (VR - a term credited to Jaron Lanier in the early 1980s), and the first to introduce a computer controlled head mounted display (HMD), the *Sword of Damocles*. He firstly described the idea of being immersed in a VE where everything, from users to objects, were generated by computer displays [Sut68]. After more than 45 years, Sutherland's words are still valid and, despite the significant progress made in the CG and VE fields, they are still a source of inspiration for many researchers.

Over the years, Sutherland and other CG pioneers' ideas have been implemented, improved and extended further. In 1978 the *Aspen Movie Map* was created at the Massachusetts Institute of Technology (MIT) by a team led by Andrew Lippman [Lip80]. The program was a crude virtual simulation of the city of Aspen, Colorado in which users could wander the streets in one of three modes: summer, winter, and crude polygonal models. The first two modes were based on photographs collected by the researchers in both seasons, while the third mode was a basic 3D model of the city. In the early 80s, multi-users VR systems introduced the paradigm of user embodiment within VE, allowing users to remotely interact in shared spaces. Recently, driven by the video games and film industries, immersive hardware devices, such has motion tracking system, large field-of-view (FoV) or head-mounted displays, and range-cameras are emerging for both commercial and domestic use.

Boosted by the demands of our modern, long-distance based society, the increased availability of immersive hardware resulted in a dramatic increase in the development of immersive collaborative virtual environment systems (ICVEs). However, the technical aspects of designing and using these new technologies are still far from being accessible to everyone. One of the biggest limitations of these mediums is the difficulty in setting-up such systems, which are usually confined to laboratories given their need for highly specialised hardware.

In contrast, in recent years portable computing research has made tremendous progress, and nowadays, with the ubiquity of video capture devices, it is very easy to record live events for real-time sharing or to form video collections. We are rapidly moving toward a world of ubiquitous video where personal networked video cameras are everywhere. With the introduction of smartphones and portable devices, such as tablet or compact cameras, owning and operating a recording device is no longer a practice left to experts and hobbyists of the field. The quality and pervasiveness of cameras on mobile devices continues to increase, while most new laptops have a built-in camera and most new smartphones and tablet-style devices have both front- and rear-mounted cameras. Rear-mounted cameras on mobile devices aim to replace or supplement the use of a normal camera, while front-mounted and laptop cameras are often used for face to face video conferencing. This allows more and more people to capture, stream and record a

Figure 1.2: *Supporting spatiality in typical immersive VMC systems often requires dedicated and expensive hardware. For instance, the Polycom RealPresence Immersive Studio [Pol11] telepresence system features 4k Ultra HD displays, 1080p video quality, a 18-foot video wall, a content touch-displays and Polycom "3D Voice" spatial audio. Image courtesy of Polycom, Inc - Press Kit.*

variety of events to such an extent that only few years ago would have been impossible. For instance, every minute 100 hours of video are uploaded to the on-line video platform YouTube [1] [Goo12], while 40% of the total calls made through Skype [Mic02], a VMC software, are video to video [2].

1.1 Research Problem

Typical immersive VMC systems and CVEs support shared "virtual spaces" in which spatiality is supported to improve the communication. Benford *et al.* define spatiality as the ability of a VMC system to support fundamental physical spatial properties such as containment, topology, distance, orientation, movement and a shared frame of reference [BGR+98]. Hence, spatiality is a critical property of most shared space systems. Indeed, such systems can be characterised according to their degree of spatiality, with the least spatial systems supporting only the fundamental spatial property of containment, and the most spatial system supporting the ability to dynamically form groups from among a larger population.

Supporting spatiality in a system is a key factor to improve interaction, which in turn can improve communication and user experience. However supporting spatiality does come with a cost. Benford *et al.* [BGR+98] argue that the associated costs with supporting high level of spatiality may be an increased implementation overhead and increasing constraints on the system interface in terms of presenting a synchronised view of the space. In particular, a high level of spatiality can be only achieved if a common context in which the action is taking place can be established. This is usually achieved by employing either expensive and dedicated hardware (e.g., [Pol10, Pol11] - see Figure 1.2) or detailed 3D models of

[1] Figures available at https://www.youtube.com/yt/press/statistics.html. Last accessed 02/10/2013.
[2] Figures available at http://www.statisticbrain.com/skype-statistics/. Last accessed 02/10/2013.

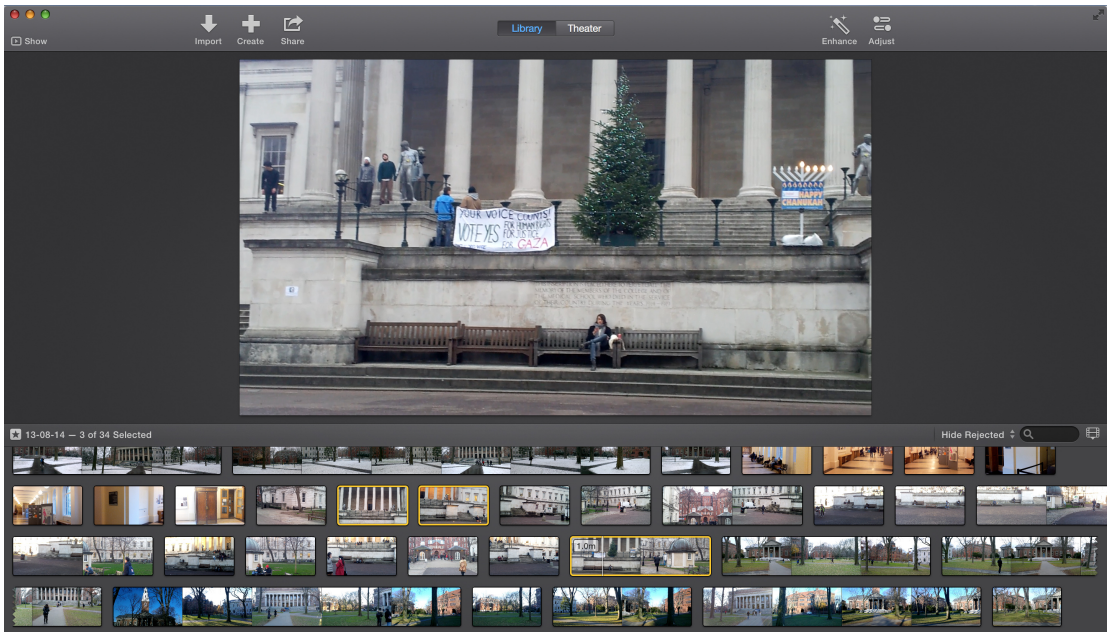


Figure 1.3: *Spatiality cannot be supported by available video browsing software. Browsing unstructured collection of videos can be hard and tedious, as spatially-related videos are not intuitively linked together. For instance, using Apple iMovie interface [App14], is hard to localise videos which are spatially related (highlighted in yellow).*

the shared environment which is often difficult to capture or unavailable. Unfortunately, such solutions tend to be laboratory based and relatively uncommon. This means that participants normally cannot access these systems without leaving their usual work or living spaces, and this constraint poses a major hindrance to the medium diffusion.

Supporting spatial awareness is not only beneficial for collaborative environments, but also while navigating and inspecting large video-collections using browsing tools, such as Apple iMovie [App14] or Windows Live Movie Maker [Mic12b]. Unfortunately, available media browsing software cannot easily convey spatial information about video collections. Browsing large, unstructured collection of videos, being them live or pre-recorded, is typically a hard and tedious task, with the abundance of visual information often being confusing and overwhelming for the user (see Figure 1.3). When the collection concerns a particular environment or place, navigating through the (sometimes redundant) videos is somehow similar to experience different parts of the remote environment. Moving through a virtual environment, and thus supporting free movement, is another important aspect of spatiality, which again requires a common context through which videos can be related. Similarly to what argued before, supporting this property comes at some cost. Researchers have tried to solve this problem with solutions that currently are far from practical for the average user, as they require detailed 3D models of the underlying spatial structure which are not always readily available or easy to obtain [BBPP10, McC07].

In general, with the existing solutions, considerable implementation and system effort may be needed to support an increasing level of spatiality (e.g., to maintain a common 3D coordinate system and to support real-time rendering with a moving viewpoint). However, the abundance of ubiquitous recording devices, and consequently the large availability of live and pre-recorded videos, opens up new

possibilities for VMC, CVE and video-collection browsing systems. When a variety of videos and images of the same location are available, a reconstruction of the environment can be created using such information, offering to the users an easily understandable visual representation. We note though, that existing algorithms cannot easily handle these vast data streams, and thus novel solutions are required.

Clearly, given multiple videos of the same location, exploiting such visual data to transmit spatio-temporal information describing the environment is not trivial. Simply showing the video feeds as they are, and outside the environmental common context, can create a confusing picture in the user's mind, with the original spatial and temporal links between cameras and videos lost in the visualisation. This is for instance the visual paradigm employed by common video browser tools, such as Apple's iMovie [App14] (see Figure 1.3). Therefore, how can these video feeds be linked spatially and temporally so that the users can easily navigate through them and feel immersed in the original environment?

1.1.1 Videos in Panoramic Context

Finding content relationships between arbitrary videos is difficult, and the field of multimedia retrieval tries to address these problems. Previous works attempted to tackle these problems by presenting a video-in-focus metaphor [NSQ12], linking the videos on a map [TKKT12] or by building a sparse 3D reconstruction of an environment an using image-based rendering for replay [McC07]. However, these solutions cannot fully capture the spatial and temporal links between videos, or are difficult to setup and operate, as they require detailed scene geometry information or dedicated hardware setup. In this thesis we propose a visual description that exploits panoramic imagery to build a visual context into which organise a network of unstructured videos or camera. Building on the concept of focus+context [CKB09], in which a subset of information is shown in full detail within a wider context of surrounding lower-density detail, we propose a visual description in which live, as well as pre-recorded, video streams are linked together using panoramic imagery as the common context. To this aim, the panorama offers to the user the wider context of surrounding lower-density detail, while the individual videos represent "focus" window which capture the details, as well as the dynamic, of the environment. We call this representation *videos in panoramic context* (or, in short, *video+contex*).

Hypothesis: The main argument that motivates our videos in panoramic context representation is that if we can automatically link sparse and heterogeneous cameras filming events that take place within the same location, then we can provide qualitative and quantitative improvements to video collection exploration and VMC systems. In particular, we believe that some aspects of spatiality, such as topology, distance, orientation, movement and a shared frame of reference, can be achieved by employing our video+context representation. We argue that by increasing capture, transmission, and display of spatial information about a remote location, VMC may be enriched, and the medium will be more able to convey a sense of space which is more similar to the one perceivable in the real world. Similarly, we believe that by automatically organising a video collection with respect to time and space, presenting this vast amount of information in its original context, users' spatio-temporal cognitive load may be eased.

We think that our representation provides an easy to setup and reliable solution to create an heterogeneous camera network that can be used for either on-line video-conferencing or off-line video

browsing. Our final goal is to obtain a visual representation that can capture the dynamics and liveliness of a place, while offering a reconstruction of the remote environment, maximising users' sense of space and, when possible, time. Our representation will have to be easily achievable by users, will have to accommodate a variety of camera types, including portable devices, and will have to scale with the numbers of cameras or videos in the collection. In addition, it will have to offer an easy to understand visual reconstruction of a place and the dynamics that happen within. In summary, our hypothesis is that the videos in panoramic context representation will be able to:

- **H1**: build a spatial and temporal graph of several videos/cameras shown together through the employment of a common, panoramic context;
- **H2**: obtain a comprehensive depiction of a remote location through dynamic videos and static imagery, improving users' spatio-temporal thinking, and consequently being beneficial for the system spatiality;
- **H3**: being achieved in a small amount of time (from few minutes to an hour, depending on the number of video streams employed), and with minimal technical intervention, relying solely on available hardware;
- **H4**: improve the sense of space and, when possible, time.

In addition, and in line with prior research carried out in the VE field, we expect that:

- **H5**: the level of immersion of a display type can be a significant factor on users spatio-temporal thinking, affecting the eventual beneficial properties offered by the video+context representation.

We wish to experimentally evaluate our hypotheses, comparing our proposed representation against existing techniques for video collection exploration and VMC.

1.2 Research Questions

The overarching goal of the research presented in this thesis has been to investigate how videos in panoramic context may be used to enhance live video-conferencing and off-line video browsing systems to improve users' spatio-temporal thinking. Additionally, the research focuses on how well this representation can replace more sophisticated visual descriptions, and if different types of display can affect users interacting with it. The research extends earlier studies in the VMC and focus+context literature, by developing two video+context systems and conducting a series of controlled experiments designed to observe the affect of videos in panoramic-contexts and display devices on users' spatio-temporal thinking.

The main experimental research, presented in Chapters 5–7, investigates various aspects of employing our proposed representation for VMC and video browsers systems. The three chapters are each concerned with two specific systems developed during this research, and document the associated experiments. Chapter 5 investigates a two-party collaborative scenario in different VMC systems, Chapter 6

explores how spatially localised video can benefit users' performance when browsing large video collection, and Chapter 7 addresses the effect of display types when coupled with video in panoramic contexts for video browsing. To summarise, the experimental work conducted during my research was guided by, and addressed, the following overall questions:

1. *Can spatially localised video be used to increase the spatial information transmitted during video mediated communication, and does this improve quality of communication between users and their spatial thinking?*

   This question is addressed by the telecommunication experiment presented in Chapter 5. The work addresses how spatially-localised video (i.e. video insets registered within a static panorama) can improve the level of spatial information transmitted during VMC, and consequently if the system spatiality is enriched and if the quality of communication between users is improved. Additionally, we investigate whether the video+context representation can substitute more sophisticated forms of remote environments description, such as fully panoramic videos.

2. *Can videos in panoramic context be used to convey spatial and temporal information describing a remote place and the dynamics within, and does this improve users' performance in tasks that require spatial and temporal thinking?*

   This question is addressed by the video browsing experiment presented in Chapter 6. The experiment addresses how multiple videos in panoramic context can improve users' spatio-temporal reasoning while browsing large collection of videos, and if the representation can be easily understood and acted upon.

3. *Measured by spatio-temporal thinking, is there an impact of display type on reasoning about events within videos in panoramic context?*

   This final question, secondary to the central focus of the research, is addressed by the video browsing experiment presented in Chapter 7. The study investigates the effect of different display type on user spatio-temporal reasoning while interacting with video+context interfaces. By keeping the visual representation constant, we vary the immersion level of the display and study if this affects users spatio-temporal understanding.

1.3 Contributions

The main contribution of this thesis is the evaluation of the use of videos in panoramic context to transmit spatial information in VMC and spatio-temporal information in video browsing systems, and how well this representation can replace more sophisticated visual descriptions of remote environments. While the work's driving motivation lies in the aspiration to enhance the affordability of collaborative virtual environments (CVEs) and the usability of video browser systems, insights into how users engage with different form of visual representations, how they respond to different display types and how these

affect collaboration and spatio-temporal reasoning are also a fundamental goal of the research. This work covers collaborative scenarios, object-based localisation experiments and VMC application and video browsing tool design and development. Additional contribution lays in the collaborative design and development of two networked immersive collaborative virtual environment systems, as technical demonstrators of the BEAMING platform (cf. Section 1.5 for an introduction). Hence, the contributions of this thesis can be classified as methodological and substantive:

1.3.1 Methodological Contributions

1. Methods to acquire, calibrate and render dynamic reconstruction of remote locations (Chapter 4). Data include imagery available from multiple camera types, including panoramic video and colour-plus-depth video, and 3D models.

2. Algorithm to handle, compress and stream colour-plus-depth videos (Chapters 4).

3. Design and development of a portable teleconferencing system (Chapters 5) and a video-collection in context browser tool (Chapters 6).

4. Experimental task designs for use in studies on spatially localised videos for VMC and browsing scenarios, and multiple display types (Chapters 5–Chapters 7).

1.3.2 Substantive Contributions

1. Research findings that address whether spatially localised videos could be used to increase the spatial information transmitted during VMC, and consequently if this can improve quality of communication and users' spatial thinking. These findings have also implications on whether more sophisticated form of visual descriptions, such as fully panoramic videos, can be replaced by spatially localised videos without degrading VMCs users' experience (Chapter 5). These findings have also implications for the design of future video VMC systems.

2. Research findings that address the impact of using videos in panoramic context to enhance users performance and spatio-temporal reasoning in tasks that require spatial and temporal thinking while interfacing with video browsing systems (Chapter 6). These findings have also implications for the design of future video browsing systems and on how well the proposed representation is perceived, understood and acted upon by users.

3. Research findings that address whether, measured by spatio-temporal thinking, display type may be an impact factor while reasoning about events within videos in panoramic context (Chapters 7). These findings have also implications for the design of future video in panoramic contexts applications.

1.3.3 Publications

Some of the content for this thesis is derived from the following publications, all appearing in peer-reviewed international conferences and journals, though here this content is significantly expanded:

- Fabrizio Pece, William Steptoe, Fabian Wanner, Simon Julier, Tim Weyrich, Jan Kautz and Anthony Steed. Panoinserts: mobile spatial teleconferencing. *In Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, 1319–1328, 2013. DOI 10.1145/2470654.2466173

- James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz and Christian Theobalt. Video collections in panoramic contexts. *In Proc. of the 26th annual ACM Symposium on User Interface Software and Technology (UIST '13)*, 131–140, 2013. DOI = 10.1145/2501988.2502013

- Anthony Steed, William Steptoe, Wole Oyekoya, Fabrizio Pece, Tim Weyrich, Jan Kautz, Doron Friedman et al. Beaming: An Asymmetric Telepresence System. *IEEE Comput. Graph. Appl.*, 32(6), November 2012. DOI = 10.1109/MCG.2012.110

- William Steptoe, Jean-Marie Normand, Oyewole Oyekoya, Fabrizio Pece, Elias Giannopoulos, Franco Tecchia, Anthony Steed et al. Acting Rehearsal in Collaborative Multimodal Mixed Reality Environments. *Presence - Teleoperators and Virtual Environments*, 21(4), 406–422, Fall 2012

- Fabrizio Pece, Jan Kautz and Tim Weyrich. Adapting standard video codecs for depth streaming. *In Proc. of the 17th Eurographics conference on Virtual Environments & Third Joint Virtual Reality (EGVE - JVRC '11)*, 59–66, 2011. DOI = 10.2312/EGVE/JVRC11/059-066

- Fabrizio Pece, James Tompkin, Hanspeter Pfister, Jan Kautz and Christian Theobalt. Device Effect on Panoramic Video+Context Tasks. *In Proc. of the Conference on Visual Media Production (CVMP '14)*

During the doctoral study for this thesis, the candidate also contributed to additional peer-reviewed publications, juried exhibitions and workshops, which are listed in Appendix A.

1.4 Scope of Thesis

This thesis is concerned with the evaluation of how well videos in panoramic context can be used to represent real environments when transmitting spatio-temporal information describing remote locations, for both VMC and video-collection browsing systems. The focus of the research is therefore not on the technologies itself, but rather on the use of the visual representation employed in the technologies to support both object-focused collaboration and spatio-temporal browsing. However, the software platforms used throughout are bespoke, and have been developed with the experimental evaluations in mind. Where appropriate, key phases of development, and system overviews are provided.

The three experiments are concerned with two different scenarios. The telecommunication experiment presented in Chapter 5 is concerned with a single user interacting with two confederates which are located remotely. Interactions is performed with different telecommunication systems, but always using non-immersive desktop displays. Chapters 6 and 7 investigate an object-localisation scenarios in which a single user is required to spatially and temporally browse a large collection of videos. While the experiment presented in Chapters 6 evaluated a variety of video browsing tools using the same non-immersive

desktop display, the user study documented in Chapters 7 investigated different display types, including non-immersive desktop display, tablet devices and HMDs. Section 7.2, Table 7.1, presents the potential design space of display scenarios. Choosing which displays to evaluate from the large number of possible configurations is not straightforward, as each display type has different properties which might not be directly comparable, and trying to normalize these conditions is difficult. Instead, we choose a systems-level approach, where we try to compare systems which would most likely be used in practice. While this makes the comparison harder, it allows us to evaluate the impact of design decision on users behaviours with system that they would commonly use. Thus, the research presented in this thesis does not consider CAVE™ immersive displays [CNSD93], which are not usually employed in widely common systems.

There is little work in the literature investigating 3D models used as context [NYH+03], and the research documented here is no exception. Neumann *et al.* proposed a system that features dynamic fusion of imagery and 3D models. However, by direct admission of the authors, the ambitious problem of aligning static imagery to 3D model presents many challenges, which if not tackled correctly, can result in a confusing visual representation. The authors note that one of the biggest challenge posed by this type of alignment lays in segmenting foreground objects (especially dynamic objects) for which no 3D models are available. Failing to do so results in foreground objects wrongly registered to the background models, obtaining in this way a confusing visual representation. A different approach to use 3D context is given by McCurdy in his telepresence system, called *RealityFlythrough* [NYH+03, McC07]. The author proposes a visual representation in which live and archived views of the scene are stitched together and situated using a 3D model of the world. However, here the 3D model is only used for rough registration, meaning that the videos and images are not accurately positioned in space and that the context does not add any spatio-temporal information to the visual description, but rather it is only used as a three-dimensional map onto which position the video feeds. In fact, McCurdy's representation heavily relies on a property of the human visual system called "closure" [McC93], which is the brain's ability to fill in gaps when given incomplete information (in this case, the absence of visual information in-between views). At the same time we note that, even though acquiring 3D models of large environments has become easier with the introduction of depth-cameras and fast stereo reconstruction algorithms, such task is still relatively hard for unskilled users. In this research we are interested in evaluating a visual description that is a) easy to understand and act upon and b) easily acquirable with any kind of device, including commonly available portable devices. As such, we decided to investigate 2.5D panoramic contexts as they ideally offer more spatio-temporal information than common video, encode the same spatial information than fully-panoramic videos and are easily acquirable. Therefore, we decide to exclude 3D contexts from our investigation in favour of panoramic imagery, but we reserve extensions to the 3D case for future work, as detailed in Section 9.3.

From an analytical standpoint, we relied on questionnaires and performance metrics, as commonly used in VE and human-computer interaction (HCI) studies, to investigate the benefits of videos in panoramic context.

1.5 BEAMING

The research presented in this thesis is conducted within an FP7 European Union funded project called BEAMING [Con10]. To contextualise this research, this section presents a brief overview of the project. However, a more comprehensive description of the BEAMING ideas and goals, and a detailed description of the design and development of two of its platform instantiations, are given in Chapters 3 and 4 respectively.

The overarching technical aim of BEAMING is to capture, transmit, and represent perceptual cues describing the activity of participants and their geographically-remote locations between sites. A *visitor* is a person that is physically absent from a destination site where a BEAMING session is taking place (i.e. the *destination*) and where other people (i.e. the *locals*) are present, but at which, through BEAMING technology, they are represented virtually. In other words, BEAMING is the process of instantaneously transporting visitors from one physical place to another so that they can interact with the locals there.

The goal of BEAMING is to provide a rich and effective telecommunications medium supporting a range of collaborative activities. This means that the visitors should achieve a high sense of presence through their virtual embodiment, feeling influential at the destination. Correspondingly, the locals should naturally respond as if the visitor is amongst them, and all parties should be able to rely on perceptual cues common to collocated communication, such as natural lines-of-sight and drawing attention via gesture. Therefore, the visitor's actions at the destination site will have physical consequences, and similarly locals' actions at the destination will have physical consequences for the visitor.

BEAMING allows remote communication between remote sites, providing a collaborative mixed-reality environment that grants symmetrical social affordance and sensory cues to all connected users whether they are locals or visitors. While remote collaboration is already possible with existing ICVE or VMC systems, the unique feature of BEAMING is that the platform abandons the symmetry of access to a shared virtual environment in which collaboration happens, and rather focuses on recreating, virtually, a real environment and having remote participants visit that virtual model. To achieve this, BEAMING supports technologically asymmetric setup that allows users to join the action regardless of their hardware. This novel ICVE system brings today's networking, computer vision (CV), CG, VR, haptics, robotics and user interface (UI) technology together to produce a new kind of virtual transportation.

With respect to BEAMING, the scope the research presented in this thesis covers all tasks that are concerned with creating and transferring a visual representation of the destination to the visitor. These include capture, representation, transmission and rendering of the destination environment and the dynamics within.

1.6 Structure

This thesis is divided into 9 chapters. Chapters 2, 3 and 4 are introductory and cover relevant research and methods, including an introduction to the BEAMING project and description of two of its instances. Chapters 5–7 present the design and findings of three studies investigating the benefits of videos in panoramic context. Chapter 8 presents a discussion on the studies findings, while Chapter 9 draws

conclusions from the findings and propose directions for continuing research.

Chapter 2 contextualises the research by expanding upon the motivation, the central problem addressed, and the general approach taken. The chapter introduces fundamental works to the fields closely related to the areas of research of this thesis, narrowing it down to the six most relevant topics. These include previous work on VMC and ICVE systems, panoramic and 2.5D video acquisition and transmission, focus+context and video+context applications, 3D reconstruction, content rendering and data fusion for large environments mapping.

Chapters 3 and *4* cover both technical and methodological aspects of the research. Firstly, Chapter 3 introduces the design and key features of BEAMING, a ICVE system which was developed (in collaboration with other researchers) throughout the course of this research, and which supported some of the experimental work conducted during the investigation. Chapter 4 documents the collaborative development of two instances of the BEAMING platform. Aspects related to the development of solutions to acquire and transmit the destination to the visitor are presented, as they are part of some of the methodological contributions of this thesis.

Chapter 5 presents the first of the three experiments which form the main empirical research contribution of this thesis. The experiment investigates the suitability of video in panoramic context for remote tasks. To support the investigation, we developed *PanoInserts*, a surrounding teleconferencing system that uses static panoramas and live videos from portable devices. The chapter documents the system architecture and development, and presents a study that compares PanoInserts with panoramic video and web-cam style video chat over an object placement scenario. Results of the study are presented and discussed.

Chapter 6 presents the second experiment, which investigated the suitability of multiple videos in panoramic context for spatio-temporal browsing of video collections. To support the investigation, we developed a second video+focus system, named *Vidicontexts*, which facilitates spatio-temporal browsing of video collections. The chapter documents the system architecture and development, and presents a study that compares Vidicontexts with existing video browsing tools over an object localisation and tracking scenario. The chapter ends with a description of the results and a related discussion.

Chapter 7 presents the final study, which concerns the effect of display type on users interacting with videos in panoramic context interfaces. To support the study, we extended Vidicontexts to work on a variety of display types which sample interesting points within the immersive displays design space. Results of the study are presented and discussed.

Chapter 8 discusses the implication of the three user studies findings, relating them back to the research questions presented in this chapter and to the overall research goal. Implications of these findings with respect to BEAMING are also discussed. Finally, *Chapter 9* draws conclusions and gives suggestions for future work.

Chapter 2

Background

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny ...'

Isaac Asimov

This chapter introduces the background to this thesis and discusses related work. The chapter aims to relate the research presented here with the literature that has shaped its motivation, the research questions it aims to address and the approaches it takes. The chapter is comprised of six main sections, which narrow down the focal areas of research to the six most relevant topics to this thesis. The first section explores long distance, human verbal communication in collocated (face-to-face) small-group interaction, with a particular focus on VMC systems and ICVEs. The second section explores the work related to video acquisition and transmission, for both the 2D and 3D cases. The third section illustrates the most relevant work on panoramic imagery acquisition and introduces work related to focus+context and video+context applications. The fourth section motivates some of the research problems by discussing the main techniques developed for 3D reconstruction, while the fifth section is focused on content rendering, with a special interest to image-based rendering. Finally, the last section explores work related to data fusion for large environments mapping.

2.1 Long-Distance Communication and Remote Collaboration

2.1.1 Video-Mediated Communication Systems

Large part of this research has been inspired and draws important consideration from work related to VMC systems. Video acquisition and rendering, an important topic of this thesis, are a crucial tasks for VMC systems, and over the last few years their development has experienced a constant rise. VMC is the most direct and accessible form of remote communication which in the last decades, following the development of our financially-conscious global society, has seen a dramatic increase in its employment. VMC has been shown to improve over audio-only communication many typical aspects of natural, face to face communication, such as the ability of giving non-verbal information or express understanding, feelings and attitudes [IT93]. For instance, the work of O'Malley *et al.* [OLA+96] shows how, when people are asked to perform a collaborative task in a VMC environment, they tend to achieve some goals

(a) *Polycom CX5000 - Image courtesy of Polycom, Inc. - Press Kit.*

(b) *Cisco Telepresence TX9000 - Image courtesy of Cisco Systems, Inc. - Official Data Sheet.*

Figure 2.1: *Example of VMC telepresence systems.*

faster and with less effort than people that can only hear each other. However, in their work the authors also stress the fact that the quality of rendering and transmission of the video significantly affect the overall communication, as when the video display or transmission is not optimal (i.e. streaming lag or rendering artefacts), the performances of the individuals dramatically drop.

Doherty-Sneddon *et al.* showed how the structure of the dialogues in VMC differs compared with dialogues obtained when users can only hear each other while performing a collaborative task [DSAO+97]. This supports the idea that the visual, non-verbal channels are extremely important for communication: the audio-only conversations have certainly more words, but these extra verbal information are replaced in a video-mediated dialogue with visual signals that can deliver the same type of content.

Given the clear advantages that VMC systems bring to remote collaboration, throughout the years many VMC systems have been developed and commercialised. These include web-cam style video-chat, recently also supporting mobile video conferencing (e.g., Skype [Mic02]), and videoconferencing tools and commercial telepresence solutions supporting high-definition video and audio (e.g., Cisco Telepresence [Cis06], LifeSize [Log03], Polycom CX5000 [Pol10] and RealPresence Immersive Studio [Pol11], BrightCom [Bri10] and Telanetix [Tel12]). Figure 2.1 shows some examples of such systems.

When referring to VMC, a closely related topic is certainly telepresence, a concept firstly introduced by Minsky in 1980 [Min80]. Telepresence is often used to describe the feeling that a human operator would experience while seeing the real world through the eyes of a machine, using his own limbs to change such world. Minsky attributed the development of the idea of a remotecontrolled system to Robert A. Heinlein's prophetic 1948 novel *Waldo* [Hei42]. In his science fiction short story, Heinlein envisioned a telepresence device through which Waldo, a man affected by profound muscle weakness, can control dozens of mechanical hands to perform his everyday life routines.

A popular application for telepresence lays in immersive videoconferencing, the highest possible level of videotelephony. Telepresence via video supports improved fidelity of both sight and sound than in traditional videoconferencing. Telepresence, then, is the feeling that most modern VMC systems try to achieve by letting its users feeling completely immersed in the remote environment they depict. To this

aim, greater technical sophistication and video rendering techniques are usually deployed to enhance the telepresence experience; such solutions include wide field of view cameras, surround videos, immersive displays and life-sized video representation of the users. For instance, Fehn *et al.* [FCSK02] propose an image-based rendering solution for 3D immersive displays. Starting from the assumption that depth perception can be reached also through brilliant quality pictures and head motion parallax, the authors present a 3-stage 3DTV system that is compatible with most of 2D displays. Similarly to the IMAX Dome system [IMA13], the authors present a way to display high-resolution imagery on large panoramic screens combining this with head-motion parallax obtained by a multiple baseline camera set-up.

Telepresence often relates to spatiality in VMC. Spatiality in mediated communication is the degree to which a system supports fundamental properties such as movement, distance, containment, topology and a shared frame of reference such as a Cartesian coordinate system [BGR+98]. A telecommunications medium supporting a high-degree of spatiality, for example shared immersive VE, presents a shared space in which all users observe from their perspective the same extents, relative positions, and orientations. Practically, this implies that spatial cues such as gestures and glances can be both performed and understood similarly to as they can be in reality. In contrast, webcam video conferencing presents disjoint portions of physical space that constrain these spatial cues, thereby hindering spatial perception and limiting gaze awareness [HRBC06]. Typical webcams feature a narrow field of view that is unsuitable for scenarios involving multiple users seated at a meeting table, or non-stationary users. While high-end commercial video telepresence systems are able to support gaze awareness provided that users remain seated, the usually static cameras do not allow for users to move around the meeting room while still remaining within the camera frame and thus visible to remote participants.

One means to foster spatial awareness in VMC, as we also demonstrate with the experimental work presented in this thesis, is to transmit a panoramic representation of a space to a remote viewer, thus overcoming limitations associated with narrow field-of-view cameras. Such cameras, often referred to as omnidirectional cameras, provide high-quality images with good sampling over the full panorama; however, they are expensive. An example teleconferencing system that utilizes an omnidirectional camera is presented by Fiala *et al.* [FGR04]. This implementation, together with commercial cameras such as the PointGrey Research Ladybug3 [Poi10b], typically assume simple cylinder, sphere or cube proxy geometry for the scene, onto which all video is projected. Alternatives, providing lower and more uneven spatial resolution, are catadioptric systems or wide angle fish-eye lenses and a single camera. Commercial systems for teleconferencing using such lenses include the aforementioned Polycom® CX5000 [Pol10]. To augment the relatively low panoramic resolution, Cutler *et al.* augment their panoramic-based VMC system with scenario-specific video insets [CRG+02]. However, contrary to the system we propose in Chapter 5, the video insets adopted by Cutler *et al.* are not spatially-related, nor are embedded in the panoramic image, but rather represent isolated video-windows, including overview or user-specific cameras. A more comprehensive discussion of work directly related to panoramic imaging is presented in Section 2.3.

Another central topic of VMC systems, and more in general of social behaviour and non-verbal

communication (NVC), is gaze [AC76]. Correspondingly, gaze awareness is a key requirement for effective VMC that has been shown to improve both task performance and sense of social presence [IKG93]. Gaze awareness, and eye contact, allow users to understand where other people are looking and eventually infer their emotions or intentions; gaze awareness may be achieved through physical alignment of cameras and displays to enable natural lines of sight. However, since VMC represents a compressed representation of 3D space, it strongly constrains the range of visual and depth cues available in normal 3D environments, making gaze awareness hard to achieve. Benford *et al.* identifies this as a key limitation of traditional video conferencing systems with regard to spatiality [BGR+98]. The authors argue that classic VMC systems do not easily support forms of spatial referencing, such as gaze direction, whereby participants can infer who is attending to whom at any moment in time from their representations. While investigations on the impact of gaze awareness in VMC systems is beyond the scope of this thesis, we note that it is an important factor for effective VMC that can simulate normal human interactions [Wil77], and we acknowledge the fact that high-fidelity video capturing and rendering is crucial to enhance the communication quality.

Finally, besides video rendering, also video streaming is an important topic for VMC. Streaming latency strongly affects the quality of VMC systems, and with the increasing improvement of video quality (i.e. HD or 2.5D videos), and correspondingly with the dramatic increase of the required bandwidth, such topic has became subject of intense research. Lamboray *et al.* [LWG05] describe how 3D video can be efficiently coded and streamed in telepresence environment by analysing the typical traffic generated by such systems. They put special emphasis on how the 3D geometry should be stored on the acquisition side, and also on the de-coupling of acquisition, processing and rendering frame rates, and audio/video synchronisation. This work extends the Blue-C system presented by Gross *et al.* [GWN+03], of which a detailed account is given in the following section. Besides bandwidth limitation, also the impact of video encoding on high definition videos and, for some systems, depth maps, has been analysed. Zia *et al.* [ZDS09] present a quantitative evaluation methodology and framework for video encoders applied to stereo data. In addition, the authors discuss a comparison of various system configurations with different performance-complexity trade-offs, giving insight on selecting the configuration suitable for a variety of telepresence applications. A more comprehensive review of work related to depth streaming is presented in Section 2.2.3.

As discussed before, while webcam-style VMC systems can be setup with minimal technical intervention, they cannot easily transmit spatial relationships between several people or objects due to cameras typically having narrow fields of view. This limitation prevents such systems to be directly used in our research, as fostering spatial awareness through videos is one of the central topic of our work. One way to overcome this hindrance is to employ fully-panoramic cameras. However, such solution requires expensive and specialised hardware, dedicated meeting rooms, or even ad-hoc multi-camera setups. In addition, users are often given only limited movement freedom. Therefore, while providing interesting inputs, VMC systems and their variations can only provide a starting point from which develop our research. Indeed, in Chapter 5 we present a portable teleconferencing system that combines web-cam style

video-chat portability with fully-panoramic video spatiality, employing a flexible and re-configurable camera setup supported by panoramic imagery.

Limitations of VMC

VMC systems are a valuable tool for enabling remote interaction through verbal and non-verbal cues. However, a major limitation of VMC is the compressed representation of 3D space, which reduces rich cues available in normal collaboration, such as depth or FoV, which consequentially limits the awareness and the interaction with users in the environment [HRBC06]. A discussion on this limitation is presented in Section 5, with an extension to fully and semi-dynamic omnidirectional video. A major drawback of VMC is indeed the inability for user to freely move in the environment without leaving the visible volume. To circumvent such limitation, some examples of novel VMC systems adapt an elaborate arrangement of cameras. Examples include movable cameras [NMK09], panoramic cameras [FGR04] or special lens arrangement [MSGF99], arrays of separate displays [SBA92], and hybrid VMC-VE systems that can operate on desktop [VWS02] and immersive displays [GWN+03].

Nevertheless, even if the aforementioned examples may alleviate some of the common spatial restrictions of VMC, users are still constrained to interact through flat, 2D windows into each other's environment. Hence, 3D space compression, restriction of movements and limited space for non-verbal communication remain the biggest limitations of VMC, as opposite to real world interaction where the spatial awareness, as well as the use of the surrounding space, are two ubiquitous features.

2.1.2 Immersive Collaborative Virtual Environment

The spatial limitations of VMC systems can be circumvented by immersive collaborative virtual environment systems (ICVE). ICVEs, similarly to VMC systems, allow collocated or remote participants to experience communication and interaction inside a rich spatial and informational context [BZD+05, MGVL02]. However, the key concept behind ICVEs is that they are shared 3D virtual worlds rather than 2D windows looking onto remote locations. Such worlds are made of computer generated spaces in which users are represented to one another in graphical forms, and can interact with each other by controlling their view points and using a variety of computer generated data [BBRG96].

Barfiled and Furness define a VE as a representation of a computer model or database which can be interactively experienced and manipulated by the users [BFI95]. However, such definition encompasses a broad set of applications, that ranges from multi-user CAVE-based VR applications to desktop-based video games. Therefore, ICVEs can be distinguished from other interaction applications through the level of *immersion* that they can achieve. Such distinction largely depends on the capabilities of the operating hardware, including displays and input devices.

Immersion has been firstly defined by Draper *et al.* as the level to which a VE can stimulate a user's sensory input channels [DKU98]. The main difference between immersive and non-immersive VEs resides in the type of interaction style and user embodiment within the environment [Ste96]. In a typical immersive VE, user's head is usually tracked so that the rendered graphics (on surrounding displays) can match as close as possible the user's viewpoint. Similarly, the user's sense of proprioception in the surrounding environment is usually maintained through the employment of bodily tracking device. In

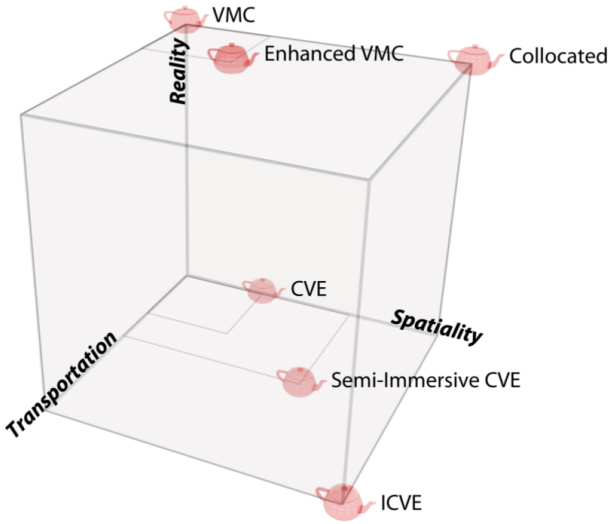


Figure 2.2: *Steptoe's visualisation [Ste10] of Benford et al. three dimensions of spatiality [BBRG96]. Note that to ensure consistent polarity of the axes, artificiality has been positively renamed as reality. Image credits: William Steptoe (Permission to reproduce this figure has been granted by the original author).*

contrast to this, non-immersive systems lack both tracking and surrounding displays, leaving the control of both user's embodiment and rendering viewpoint to standard input devices. This obviously limits the level of immersion that can be reached in such systems. Slater summarises the difference between immersive and non-immersive system by stating that, in an immersive system it is in principle possible to simulate a non-immersive system experience, but non vice-versa [Sla09].

Presence, often also referred to as *place illusion* (PI) [Sla09], is another key features of ICVEs. Presence is best defined as the user's feeling of "being there", in a computer- generated environment [SUS94]. Slater states that immersive VR systems are characterised by the "sensory-motor contingencies" (SCs) that they support, referring to the actions that a user can perform to perceive some aspects of the VE [Sla09]. For instance typical SCs of immersive VE is the ability to bend down a table to see what is underneath, or simply rotate head and eye to change gaze direction. Full immersion is crucial to maintain the illusion of reality and consequently, presence. Realistic feedbacks from the VE, such as verbal of non-verbal avatars responses, are extremely important to avoid breaking the place illusion, as noted by Pan and Slater [PS07].

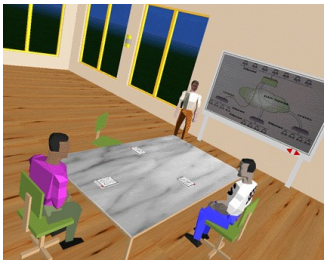
Another key features of ICVEs is the level of *spatiality* achievable through them. We have already discussed how spatiality can be fostered in VMC systems, albeit the employment of costly and intrusive hardware. Steed *et al.* state that ICVEs are particularly suited to highly spatial and interactive tasks, as collaboration is intuitive [SSH+03]. Benford *et al.* suggest that spatiality can be represented through two fundamental dimensions, *transportation* (which is analogous to presence), and *artificiality*, which describes the extent to which the shared space is either artificial or based on physical world [BBRG96]. In addition, the authors identify spatiality as the main dimension in which system can be ranked, and that represents the degree to which a system support key spatial properties such as distances or topology. In their seminal work on spatiality, Benford *et al.* presented two graphs that ranked various shared space

systems with respect to dimensions of transportation and artificiality and spatiality, respectively. Steptoe revisited this visualisation in [Ste10], and created the more intuitive graph which is reported in Figure 2.2. The graph clearly explains the key spatial differences between VMC and ICVE systems, highlighting the main features that differentiate the two shared space typologies.

Researchers have tried to quantify the effect of display devices on user on user immersion and thus performance. Several experiments [BDR+02, TKBMW12] have compared immersive displays, such as CAVEs or HMDs, to traditional displays. The early work of Slater focuses on how immersive displays might afford users a greater sense of presence [SU93, SLU+96, SSA+01], and his studies discover that immersion can lead to increased performance in 3D spatial tasks. Bowman al. [BDR+02] investigated human behaviour and performance when using a HMD and a CAVE, discovering that HMD users are significantly more likely than CAVE users to use natural rotation in a VE. Other research measures the relationship between display type and spatial reasoning [PSP93, RSPB05]. They find that, along with HMDs, large projection screen systems can also offer a greater sense of immersion which may lead to better performance. Mizell et al. find that immersive displays can better convey the sense of space than desktop displays [MJSS02]. Patrick et al. [PCS+00] compare various displays which occupy comparable visual angles, and find that, while users performed significantly worse in forming cognitive maps on a desktop monitor, users performed no differently using a head-mounted display or a large projection display. Similarly, Tan et al. [TGSP03] studied the effect of large projected wall displays, and suggest that large displays afford a greater sense of presence, leading to better performance.

All the studies presented so far focus on comparing different display types while keeping the rendered content unmodified. Polys et al. reverse this approach and investigate the effect of software field of view (FoV) on user performance [PKB05]. The authors find that, for similar displays, higher FoVs benefit search tasks by showing more of a scene in the periphery, but worsen accuracy in the comparison task by distorting a scene object's spatial location.

Immersion, presence and spatiality are then the main features that differentiate ICVEs from VMC systems, as clearly shown in Figure 2.2. While VMC systems are likely to remain superior in terms of presenting truthful appearance of the users, ICVEs can offer unified shared spaces in which remote actions are propagated to each user, improving the interaction and communication between fellow participants [Ste10]. When relating back to the work presented in this thesis, ICVEs offer interesting inputs from which the research can develop. However, is important to note that these systems present major limitations that prevent them for being used directly in this thesis. While ICVEs are designed to support group collaboration, and feature multiple cameras and immersive displays to achieve gaze awareness and a sense of space, they require equipment to be installed in a dedicated meeting room and also assume that the 3D structure of the shared environment is known a priori and users are fully tracked. Hence, ICVE systems are usually expensive, lack portability, and require large technical interventions to be installed. Such constraints go against our goal to enable easy-to-access remote collaboration, and in general they contradict the hypothesis introduced in Section 1.1.



(a) *A virtual environment conference with three people simultaneously connected to the DIVE system [AFH+97]. Images publicly available [Fah97].*

(b) *A person interact with remote colleague in the Office of the Future [RWC+98].*

(c) *An user of the Blue-C system [GWN+03]. Images publicly available [Moe].*

Figure 2.3: *Examples of ICVEs systems.*

Examples and Limitations of ICVEs

The last few years have witnessed a dramatic growth in the number, as well as in the variety, of distributed VE systems, of which Meehan gave an interesting survey in 1999 [Mee99]. Hence, ICVEs in which users are embodied by avatar representations (i.e. graphical humanoids) have rapidly increased in prevalence and popularity as an emerging form of visual remote interaction [DYNM06, SOM+09].

Some of the most notable ICVEs are Greenhalgh and Benford's MASSIVE [GB95], Wolff *et al.* 's EyeCVE [WRM+08] (of which an investigation of its use is given in Steptoe *et al.* [SSRR10]), Frecon's Distributed Interactive Virtual Environment (DIVE - Figure 2.3(a)) [Fre03, AFH+97], Raskar *et al.* 's Office of the Future (Figure 2.3(b)) and Gross *et al.* 's Blue-C [GWN+03] (Figure 2.3(c)).

Blue-C, an immersive projection and 3D video acquisition environment for telepresence and collaboration, combines simultaneous acquisition of multiple life video streams with advanced 3D projection technology in a CAVE-like environment. The peculiar feature of Blue-C is the use of three rectangular projection screens (based on active stereo) which are built from glass panels containing liquid crystal layers. These screens can be switched from a whitish opaque state to a transparent state allowing the video cameras to capture through the walls. The projectors are synchronously shuttered along with the screens, the stereo glasses, active illumination devices, and the acquisition hardware. From multiple video streams, Blue-C computes a 3D video representation of the user in real-time and then streams it over the network. As Blue-C includes some of BEAMING's key features, it can be considered as its predecessor, and certainly a source of inspiration for its development. However, given the variety of technologies employed, the types of visual representation adopted and its technical asymmetric setup, BEAMING presents a novel type of ICVE and thus differs from Blue-C (and in general from all the other ICVE platforms mentioned before).

A recent example of collaborative VE is given by Maimone and Fuchs [MF11]. The system shares some of the ideas proposed in BEAMING, especially regarding the acquisition side, offering a room-sized telepresence system with fully dynamic real-time 3D scene capture and continuous-viewpoint head-tracked display on a life-sized tiled display wall. However, similarly to the VMC systems introduced earlier in this section, the platform only provides a virtual window on the remote locations rather than a shared space in which users can interact. Maimone *et al.* [MYD+13] present a variation of

this system by replacing the display wall with an optical see-through head-worn display, increasing the spatiality and immersion of the original platform.

As mentioned before, the distributed VEs examples mentioned above are strongly constrained by either their technical setup or expensive and encumbering hardware. In our research we aim to overcome these two major limitations, enabling easy-to-access remote collaboration. This is particularly true for the BEAMING system, which therefore differentiates itself from the aforementioned examples for several reasons. Unique features of BEAMING can be found in its hardware heterogeneity, different forms of visual representations, and in the employment of haptics and spatial-audio to enhance the communication experience. Additionally, BEAMING aims to be a portable solution such that people can use it from anywhere in the world and with a variety of portable devices. With this respect, the aforementioned ICVEs are certainly not portable.

2.2 Video Acquisition and Transmission

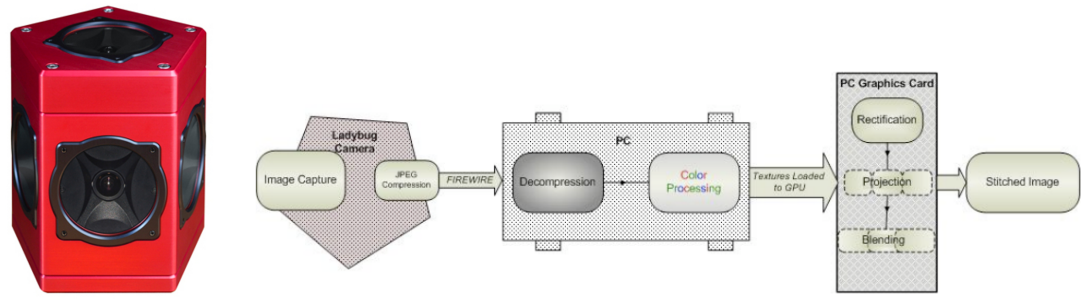
The work presented in previous sections demonstrates that video acquisition, and most importantly video quality, are extremely important for long-distance communication and collaborative environments. Therefore, novel techniques to capture video streams at higher resolution have been developed in the last few years, and this resulted in the creation of better compression algorithms that reduce the streaming bandwidth while preserving image quality. However, recently novel camera types, such as depth or omnidirectional cameras, have been introduced in the market, opening up new research opportunities in the fields of video compression and streaming.

Due to the BEAMING's technical hybrid nature, part of my research focused on integrating heterogeneous cameras into a single framework that can offer a rich and detailed description of an environment. When developing any form of ICVE systems, data acquisition must be integrated with efficient data streaming. While standard video streaming has been largely investigated by the research community, depth-enabled transmission still remains an open problem with few, ad-hoc solutions. Typical depth data consists of large collections of three dimensional points that might be associated with additional information such as distance, colour or normals. Additionally, they can be created at high rate and therefore occupy a significant amount of memory resources. Once depth-maps have to be stored or transmitted over rate-limited communication channels, methods for compressing this kind of data become highly interesting. However, in the context o ICVEs, such methods must also work at interactive rates and with minimal latency.

Hence the remainder of this section is organised in two parts. The first part presents relevant work related to panoramic and video-plus-depth acquisition, while the second part introduces work directly related to depth streaming.

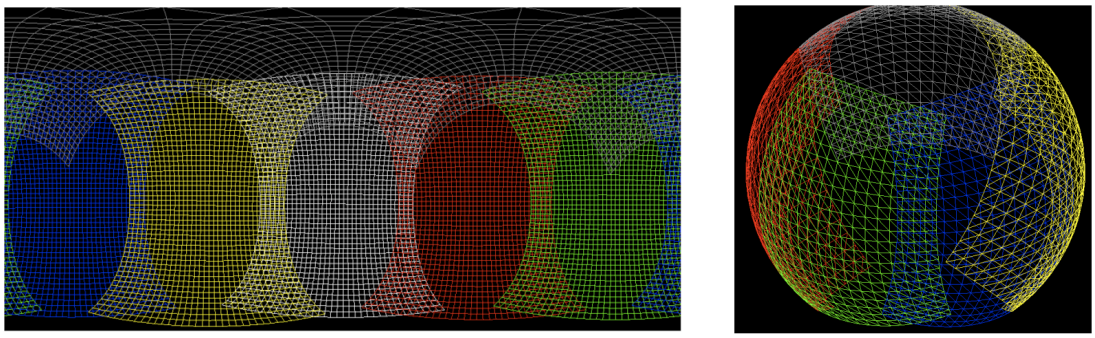
2.2.1 Omnidirectional Videos: Panoramic Cameras

The automatic construction of large, high-quality panoramas from hand-held photographs or video is a research topic that has been extensively explored in the last few years. As this section will only cover omnidirectional video and related hardware, we direct the reader to Section 2.3 for an extensive



(a) *A Ladybug3 camera.* **(b)** *Ladybug3's panoramic imaging pipeline.*

Figure 2.4: *The Ladybug3 hardware and high-level imaging pipeline. Images available from [Poi10b, Poi11].*



(a) *2D polygon mesh.* **(b)** *3D polygon mesh.*

Figure 2.5: *Polygon meshes employed in the Ladybug3's stitching technique. Images available from [Poi11].*

review of work related to panoramic imaging construction. Typically, panoramic imagery is computed by stitching together several views of a scene captured under different view points. The stitching technique, though, is not unique, and in fact two main methods are commonly used: stitching based on geometric transformation and stitching based on image content.

The first method is the one employed by a widely used panoramic camera, the Point Grey Research Ladybug3 [Poi10b] (Figure 2.4(a)). In the last few years, Point Grey Research emerged as one of the leading companies for a range of different off-the-shelf vision solutions. In particular, the company has developed several hardware and software solutions for omnidirectional video known as Ladybug. The Ladybug3 camera is capable of performing all the image acquisition, processing, stitching and correction necessary to integrate multiple camera images into full-resolution digital panoramic videos. A key feature of the camera is that the whole process is extremely fast, allowing real time acquisition of panoramic footage. The diagram in Figure 2.4(b) provides a high-level overview of the steps required to produce a panoramic image from the camera's raw images output.

The Ladybug3's capture pipeline starts with the acquisition of six different images, which are then converted from analog to digital raw (Bayer tiled) format, compressed into JPEG format and thus sent onto the fire-wire bus. When received on the PC, the images are decompressed to raw format, and colour processing is applied. During this step, the raw Bayer tiled images are interpolated to create a full array

(a) *The panoramic camera.* **(b)** *The final output of the camera.*

Figure 2.6: *The panoramic camera (Left) and its resulting frame (Right) presented by Majumder et al. [MSGF99]. Image courtesy of Majumder et al. .*

of RGB images using specific algorithms (i.e. nearest neighbour, edge sensing, high quality linear). If specified, falloff correction is applied to remove the vignetting effect of the lens. Following colour processing, the six images are loaded onto the graphics card of the PC for rectification, blending and stitching. This last step is responsible for the panorama creation: the PC's graphics card maps the image textures onto polygon meshes that reflects a panoramic view. Figure 2.5(a) shows the stitching region for a panoramic view. The coordinates of the polygon meshes are calculated based on calibration data, which specifies how to rectify, rotate and translate images. Because the textures also contain the pixels' alpha values, image blending results tend to be smooth. The stitching process is completely performed on the graphics card: image textures are firstly transferred to the card where a single, stitched image is produced without consuming any CPU resource. As it is clear from Figure 2.5(b), the same panoramic texture can be easily mapped to a sphere to reproduce the 360° spherical view of the camera. Section 4.1 illustrates how to implement and render such mapping.

Ladybug3's stitching algorithm is based on geometric transformations, allowing the camera for real time stitching of panoramic images. Unfortunately, this method is error prone, and as a result the Ladybug3 stitching often contains some visible artefacts. Firstly, as the six cameras that form the entire unit are arranged in a pentagon shape with an additional unit placed the top (Figure 2.4(a)), the various images that form the panorama are not obtained from the same viewpoint. Second, to perfectly stitch images together, the depth of the physical scene is required. However, since the Ladybug3 cannot perform range estimation, during the stitching process all points in the scene are assumed to be at the same radius from the camera, effectively forcing the real world to be mapped to a sphere surrounding the camera. Such mapping results in a compression of objects with a depth different from the assumed radius, and consequently this introduces some visible artefacts.

Majumder *et al.* [MSGF99] introduce an omnidirectional camera whose stitching technique, unlike the Ladybug3, is based on a careful mirror alignment and geometric transformations. The device comprises of six camera arranged as a cluster around a set of mirrors (Figure2.6(a)). Each camera in the cluster sees a trapezoidal mirror from which the world space is re-projected. The trapezoidal region of interest in each image is geometrically and photometrically registered with its neighbours to construct a

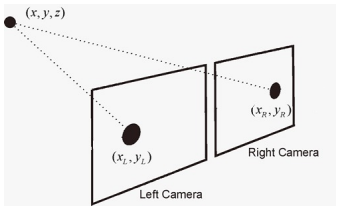
panoramic image of the world from a common centre of projection (COP). While the camera is capable of achieving real-time panorama acquisition, its horizontal FoV does not capture the full surrounding space, but rather it is limited to approximately 180° (Figure 2.6(b)). A similar approach has been investigated by Weissing *et al.* [WSEK12] with their *OMNICAM*, a scalable camera system which can be equipped with up to twelve HD cameras for panoramic capturing. Similarly to Majumder *et al.* , their panorama stitching approach is based on a flexible, mirror-based multi-camera rig that uses multiple HD cameras to capture high-resolution video panoramas. In its current implementation *OMNICAM* uses six HD cameras suitable to shoot 180° panoramas. Using mirror to capture omnidirectional video has also been explored by Fiala *et al.* [FGR04] with their video system composed of a digital video colour camera fitted with a panoramic lens mirror assembly. The camera captures a digital color video stream of which an annular region of 800 pixels diameter contains a 360° view of the environment. The authors employ this camera in a VMC system, and couple the device with a microphone array which is used to localise candidate talkers to efficiently select and transmit only a portion of the panorama to the other end of the system. As such, the view directly acquired by the camera, or cropped region thereof, is unnatural to present to a human viewer and thus has to be converted to the image that would have been seen by a traditional camera. A first transformation warps the useful pixels in the raw image into a standardized panorama accounting for all device specific parameters, such as focal length and radial profile. A second transformation produces a final image with correct perspective which is then streamed to the other end of the system.

Limitations

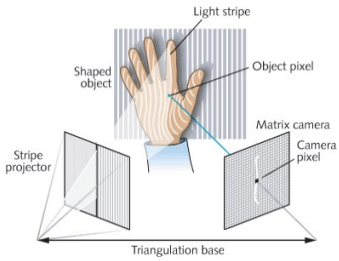
In Section 2.3 we will discuss other type of stitching techniques – the ones based on matching image content. This type of algorithms cannot easily reach interactive performances, the geometrical stitching techniques illustrated in this section remain the best solutions for live panoramic video acquisition. However, it is important to note that the works present in the last part of the previous section [MSGF99, FGR04, WSEK12] represent ad-hoc solutions which a) require careful camera or mirror alignment, limiting the portability of the technique and b) typically capture only a portion of the full panoramic view. On the contrary, the on-hardware solution offered by the Ladybug3 camera is relatively portable and is capable of capturing full 360° FoV panoramic images. Hence, it offers the best option for real-time omnidirectional video acquisition and, consequently, for the BEAMING platform.

2.2.2 2.5D Videos: Depth Cameras

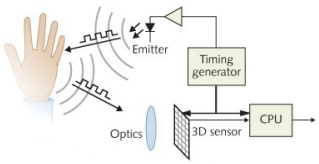
There is a vast amount of applications that benefit from or even require geometric information acquired from real environments, such as virtual and augmented environments or human-computer interaction. As such, depth cameras are a core component of many machine vision systems, and therefore a vast range of range technologies is available nowadays. The notion of depth camera, also known as *range imaging*, subsumes contact free techniques for acquiring per-pixel distance information with respect to a scene. Even though, a single acquired frame formally results in 2.5D information, a large number of applications use a set or series of these 2.5D data sets to achieve full 3D information; therefore we denote the data delivered by range sensing systems as 3D data. With the term 2.5D we refer to a particular type



(a) *The stereo vision principle. The same point is identified in two different views and its 3D location is retrieved through triangulation.*



(b) *The structured light principle. A light patter is projected in the scene by a light emitter. Subsequently a camera sensor captures the deformed pattern and depth variation is computed by analysing the way the projected pattern deforms when striking an object surface in the scene. Image courtesy of Softkinetic-Optrima [Sof11].*



(c) *The Time-of-Flight principle. The system measures return time by observing the phase shift between returned and emitted pulses. Image courtesy of Softkinetic-Optrima [Sof11].*

Figure 2.7: *Three depth-camera technology's principles.*

of video source that holds for each pixel, besides colour information, also a depth value. While some range cameras offer both colour and depth information (i.e. stereo and structured light cameras), a subset of them needs to be augmented with normal colour sensors to create 2.5D videos. This is the case of most time-of-flight (ToF) hardware.

As being able to process depth information is becoming increasingly appealing for a variety of research communities - vision, graphics, telecommunications but also natural user interfaces, to name a few - the last few years have seen a booming interest in the development of techniques to perform range estimation. In the next sections we will discuss the two basic principles for range imaging, namely ToF and triangulation-based methods, with the latter including stereo and structured light (SL) techniques .

Stereo Cameras

Stereo vision is an imaging technique that can provide depth measurements in unstructured and dynamic environments. Stereo cameras usually provide depth measurements at standard video frame-rate (i.e. 30 Hz). However, such frame-rate is usually only achievable at lower resolution or with not so accurate depth extraction algorithms. The foundation of stereo vision traces back to 3D perception in human vision and is based on triangulation of rays from multiple viewpoints. Similarly to the human brain that perceives depth by means of two dissimilar pictures [Whe38, SSG00], stereo cameras use two or more views displaced horizontally from one another to compute a depth value for each pixel in an image (see Figure 2.7(a) for a graphical overview). An analysis of decrease of accuracy in passive stereoscopic vision shows that to obtain absolute depth estimates useful for robotics it is necessary to measure all mechanical parameters with extremely high precision [VT86].

Broadly speaking, stereo cameras can estimate depth by comparing multiple images. To obtain a meaningful image comparison though, each view needs to be transformed as if it was observed from a common projective camera. There are several ways to achieve this, and the most common one is projecting one camera to the other. Once this is done, the parallax between the views makes the shift of the cameras (i.e. disparity) clear. In practice this process, usually referred to as the projective reconstruc-

tion, requires some additional steps. These are distortion removal, image rectification (i.e. projection of the images into a common plane), and disparity computation and inversion to identify the real depth measure [ZF92, ZDFL95, Zha97]. An interesting survey on different methods to perform projective reconstruction is given by Rothwell *et al.* [RFC97].

With regard to the stereo hardware used for this research (i.e. PointGrey Bumblebee XB3 [Poi10a]), the method used by the camera slightly differs from the one described above, as the unit employs three views rather than two. Each pixel in the three images collects light that reaches the camera along a ray. If a feature in the world can be identified as a pixel location in an image, then this feature lies on the ray associated with that pixel. Using multiple cameras means that multiple rays can be employed, and their intersection results in the 3D location of the feature, and hence its depth. This process is usually referred to as point triangulation. Therefore, the problems to solve are now twofold: identify the feature correspondences and calibrate the views to perform the point triangulations. For the first task, there are many solutions available in literature, see Weng *et al.* [WAH92], Tuytelaars and Mikolajczyk [TM08] and Szeliski [Sze06] for detailed surveys. Bumblebee cameras though use the sum of absolute difference (SAD) correlation algorithm to identify matching features. For the triangulation, we firstly need to perform projective reconstruction to allow view comparison (that in this case can also be identified as the image rectification or camera calibration), and then compute the matching pixel-ray intersections. Bumblebee cameras use a stereo rig method [ZLF96, Zha95] to perform camera calibration.

An alternative to the PointGrey Bumblebee XB3 is offered by Videre Design with their range of stereo solutions [des10]. The company offers a fixed baseline, as well as a variable baseline stereo camera that is capable to acquire depth information at medium and VGA resolution. The underlying stereo algorithm is similar to the one employed by the Bumblebee. Urmson showed that the Videre Design cameras, when operated at VGA resolution, can obtain similar stereo range results to the ones produced by the trinocular Bumblebee camera [Urm00]. However, as a result of their small memory footprint algorithm for depth estimation, the solution offered by Videre Design is up to six times faster at stereo processing than the PointGrey hardware.

Structured Light Cameras

SL cameras, such as the first version of the Microsoft Kinect [Mic12a], employ a depth sensor to estimate depth values from real scenes by continuously projecting an infra-red structured light pattern. SL solutions usually employ an infra-red laser projector combined with a monochrome CMOS sensor which captures depth variation by analysing the way the projected pattern deforms when striking an object surface in the scene (Figure 2.7(b) illustrates this principle). This is the same principle that is used for SL 3D scanners. Historically, structured light scanner have always operated at low frame-rate. However recently Liu *et al.* [LWL+10] have made a major speed breakthroughs in 3D laser scanning by introducing a technique that can reach data processing at 120 Hz. By utilizing the binary defocusing technique, Zhang and Huangeven [ZVDWO10] have shown that even higher working frame-rate, well over 500 Hz, can be achieved without loss of precision.

There are different SL approaches in literature: Zhang and Peisen [ZH04] propose a real-time SL

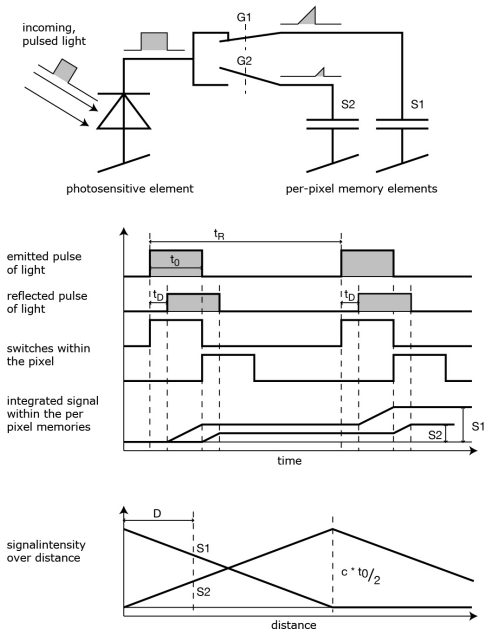


Figure 2.8: *Diagrams illustrating the principle of a ToF camera with analog timing. Image [Ogg10].*

system which runs on specialised hardware. The authors present a real-time scanner that uses digital fringe projection and phase-shifting technique to capture, reconstruct, and render high-density details of dynamically deformable objects, such as facial expressions, at 40 Hz. Scharstein and Szeliski [SS03] introduce high resolution depth maps of complex scenes that are generated using multiple SL projectors. Zhang *et al.* [ZCS02] propose a variation of classic structured light algorithms that uses a pattern of stripes of alternating colours to match observed edges in the scene and recreate 3D shapes; similarly to this, Fechteler *et al.* [FER07] propose a fast and high resolution 3D scanner that recreates textured 3D shapes from just two images. Hall-Holt and Rusinkiewicz [HHR01] introduce a SL method to obtain real-time structured light range scanning based on a new set of illumination patterns. Such patterns are based on coding the boundaries between projected stripes. The stripe boundary codes allow range scanning of moving objects at 60 Hz with 100 µm accuracy over a 10 cm working volume. The system uses a standard video camera and digital light processing projector (DLP) to produce dense range images.

Time of Flight Cameras

ToF cameras, such as the PMD[vision] CamCube3.0 [PMD09], CSEM SwissRanger [CSE06] or the SoftKinetic DepthSense [Sof11], provide robust depth data of real world scenes at normal video frame rates. Unfortunately, currently available cameras offer limited resolution with depth data sometime influenced by random error which makes them inappropriate for high-quality 3D scanning [STDT08].

An excellent explanation of the ToF principle is given by Gokturk *et al.* [GYB04]. ToF cameras use light pulses: the illumination is switched on for a very short time, the resulting light pulse illuminates the scene and is reflected by the objects. The camera lens collects the reflected light and mirrors it onto the sensor plane. Depending on the distance, the incoming light is delayed, and the sensor can compute the physical distance accordingly (see Figure 2.7(c) for a graphical overview). Figure 2.8 illustrates the

principle of a time-of-flight camera (please note that for the case of PMD camera, contrary to the figure, the function used as modulation signal is a sine-wave). In the diagram the pixel, which consists of a photo diode, uses two switches ($G1$ and $G2$) and two memory elements ($S1$ and $S2$) to convert the light to distance value. The switches are controlled by a pulse with the same length as the light pulse, where the control signal of switch $G2$ is delayed by exactly the pulse width. Depending on the delay, only part of the light pulse is sampled through $G1$ in $S1$, the other part is stored in $S2$. Depending on the distance, the ratio between $S1$ and $S2$ changes as illustrated in the drawing.

A large number of ToF related work has emerged in the last years. Ulrich *et al.* [USRO02] give an estimate on ToF accuracy by analysing the results of different time-of-flight units while acquiring 3D data of naturally reflecting objects within a wide FoV and with ranges of up to 1000 meters. Lang and Pai [LP99] present a Bayesian method to estimate distance and surface normals by using a ToF camera. The method provides more accurate estimates of range for dark surfaces that are usually difficult to measure. Finally, similarly to the structured light case, Jarvis [Jar83] provides a laser scanner that exploits the ToF principle for object reconstruction.

Limitations

Each of the three depth sensing techniques introduced in this section suffers from specific limitations, which often hinder their applicability to general scenarios. Stereo cameras, while commonly supporting medium to high resolution frames, often struggle in obtaining dense depth map from real-life scenes. Stereo algorithms cannot work in non-textured or shiny areas, making the employment of such solution for dense scene acquisition problematic. Additionally, the quality of the depth information retrieved by a stereo camera is usually inversely proportional to the speed of the algorithm employed. Therefore, capturing 2.5D videos at interactive rates with a stereo camera usually results in poor depth estimation.

Similarly to stereo cameras, SL techniques can be affected by scene lighting and objects' material, and therefore are often constrained to work indoor. Translucent objects are hard, if not impossible, to capture with SL cameras. Additionally, depth samples located far away from the camera are typically erroneously retrieved, with the intensity of the noisy increasing with the distance from the sensor.

Also ToF sensors suffer from several limitations, preventing them from being a reliable source of depth information. ToF cameras produce noisy depth data at low resolution or in scene with fast motion, and they do not deliver spatio-temporal correspondences which are essential for many advanced video effects. Moreover, ToF solutions usually offer less scene information than stereo or SL cameras, and can be subject to light conditions and objects material. A solution to these limitations has been proposed by Richardt *et al.* [RSD+12]. The authors illustrates the steps necessary to construct a computational video camera that is capable of producing spatio-temporally coherent colour-plus-depth videos at interactive frame rates. While in its current state ToF solutions cannot be directly employed for large scale reconstruction, it is interesting to notice that the next generation Microsoft gaming console, the *Xbox One*, will feature a ToF sensor to enable depth sensing and gesture recognition [Kni13]. This, similarly to the SL explosion fuelled by the introduction of the first generation of Kinect cameras, may push the boundaries of ToF research further, and consequentially produce cameras with better and more reliable data. Never-

theless, for the time being SL cameras offer a fast and reliable solution for dense scene acquisition. For this reasons, they are well-suited for the research presented in this thesis.

2.2.3 Depth Streaming

The reader should now be aware that in the last few years depth acquisition has become a popular topic of research. This has reflected in an larger availability of depth cameras that allow direct acquisition of scenes' depth information. However, while there is a large number of applications that can take advantage of this, new problems are introduced. For instance, streaming depth-video sources is a non-trivial task, typically due to the type of data employed (i.e. 16-bits per depth or higher) and, consequently, the required bandwidth. Hence, depth streaming is a novel problem with only few, ad-hoc solutions. While some work has been done to develop specific depth codecs, the same cannot be said for solutions adapting depth maps to conventional video streaming. Recently, an open and royalty free video compression standard, named VP9, has been developed by Google [GH13]. VP9 is the successor of VP8 video codec, and it is the first open codec to officially support depth encoding and decoding.

Depth streaming is a central topic in free viewpoint video (FVV) and 3D television (3DTV) [KAF+07] applications. An interesting overview of suitable technology for such applications is given by Smolic and Kauff [SK05]. A popular format for 3DTV uses a conventional monoscopic colour video and an associated per pixel depth image corresponding to a single, central viewing position. This format, named "video-plus-depth", has been adopted by the ATTEST system [RdBF+02], one of the first project that could demonstrate the feasibility of a 3DTV processing chain. By employing such format, the ATTEST system is able to obtain backwards compatibility to existing 2D services for digital video broadcast, efficient compression capabilities and a high adaptability to 3D display properties and viewing conditions [Feh04]. While the monoscopic video stream is encoded with the standard MPEG video coding, the auxiliary depth information is compressed by using an adapted version of the H.264/AVC standard [MWS06]. As a first effort towards standardisation of technologies for 3DTV and FVV applications, a new standard addressing algorithms for multi-view video (MVV) data compression — Multi-view Video Coding — has been developed by the Joint Video Team (JVT) of VCEG and MPEG [IMYV07]; however, multi-view video coding is intended to encode stereoscopic (i.e. two views) images by adapting the H.264 codec [MBX+06], and as such it does not lend itself for direct depth encoding.

Video codecs are typically optimised for images and human perception, and thus a naïve adaptation of such codecs to the depth case would not sufficient. Instead, Merkle *et al.* [MMS+09] acknowledge the need of special solutions to enable video codecs to depth compression. In their work, the authors present a different depth-optimised encoding for adaptive pixel blocks that are separated by a single edge, and assign to such block a constant or linear depth approximation. Pajak *et al.* [PHE+11] present an automatic solution for efficient streaming of frames rendered from a dynamic 3D model. The proposed algorithm is based on an efficient scheme that relies only inter frame prediction, ignoring future frame predictions. Maitre and Do [MD08] present a different approach based on joint colour/depth compression. The authors exploit the strong correlation between colour and depth to develop an ad-hoc codec that relies on a shape-adaptive wavelet transform and an explicit representation of the locations of major

depth edges. However, this solution is limited by its semi-automatic approach. Also region-of-interest specifications and depth-value redistribution can improve depth compression and transmission quality, as showed by Krishnamurthy *et al.* [CSSH04].

Interesting solutions for depth compression have also been developed for telepresence and VMC systems. Lamboray *et al.* [LWG04] propose a communication framework for distributed real-time 3D video rendering and reconstruction. They introduce several encoding techniques and analyse their behaviour with respect to resolution, bandwidth and inter-frame jitter. Würmlin *et al.* [WLG04] propose a point-based system for real-time 3D reconstruction, rendering and streaming. As their system operates on arbitrary point clouds, no object shape assumptions are made, and topological changes are handled efficiently. Recently, Kammerl *et al.* [KBR+12] introduced an entropy based, lossy point cloud compression. Their compression method, based on octree structures, is well-suited for sparse 3D information. By exploiting octrees structures properties, the authors enable fast spatial decomposition, efficient neighbour search, and compression of temporal redundancy of point cloud streams. However to achieve at interactive rates, the method requires strong compression, thus dramatically down-sampling the transmitted cloud.

Limitations

Most of the solutions presented in this section lacks generality, as they rely on ad-hoc alterations of existing video codecs and are strongly tied to specific applications. As such, they cannot be directly integrated into existing streaming pipelines. An exception to this is the work of Kammerl *et al.* [KBR+12], which presents a general point-cloud compression solution. However, such solution cannot be directly leveraged for real-time applications, given its heavy computational load. For these reasons, during our research we investigated the possibility to obtain a general solution to stream 2.5D videos in real-time while using existing video-codec implementations. This resulted in the development of our depth streaming solution [PKW11], which allows the employment of unmodified video codecs for efficient depth encoding and decoding, which we introduce in Chapter 4.1.1.

2.3 Panoramic Imaging and Videos in Context Applications

The background work introduced in previous sections has explored several aspects of video acquisition and transmission and their implication to telepresence and VMC. However, a large part of this research investigates the suitability of videos in panoramic context for both immersive telecommunication systems and spatio-temporal exploration of remote locations. Therefore, a large body of work that has inspired this research can be found in the efforts of the panoramic imaging construction and acquisition community. Additionally, the research presented in this thesis is inspired and draws important consideration from work related to the general problem of spatio-temporal media exploration and video in context applications.

2.3.1 Panoramic Imaging

Panoramic imagery is a photography technique which, by using specialized equipment or software, captures images with elongated fields of view which covers a FoV approximating, or greater than, that of

(a) *160° panorama showing San Francisco in ruins following the 1906 earthquake – George Lawrence, 1906.*

(b) *360° panorama of Philadelphia city center – unknown author, 1913.*

Figure 2.9: *Early panoramic images.*

the human eye – about 160° by 75°. This generally means that a panoramic frame has an aspect ratio of 2 : 1 or larger, the image being at least twice as wide as it is high. The origin of panoramic imagery, and even cameras, can be dated back to the mid 19th century. However, panoramic photography become popular only following the invention of flexible film in 1888, with dozens of panoramic cameras being marketed. Figure 2.9 shows two examples of early panoramic imagery. Panoramas make an attractive context for videoconferencing and video browsing applications as they provide wide or omni-directional views of an environment in a single image. It is a testament of the success of these techniques that there are panoramas publicly available on-line for hundreds of thousands of places, through mapping portals such as *Google Street View* [Goo07] or photography platforms like *Panoramio* [Goo08b].

There are many established methods to construct panoramas, which can be broadly classified in two main classes. The first class is based on special hardware, and includes solutions based on well-calibrated cameras [FGR04] or special camera and mirror arrangements [MSGF99, WSEK12] (see Section 2.2.1 for more details). The second class is based on image-based algorithms, and includes registration of multiple videos [AZP+05, SPS05] or stitching of overlapping still images [Sze94, SS97, BK01, BL07, Sze10]. While the first class of methods provides a fast and reliable solution to construct panoramas, its accessibility is limited by the high costs of the hardware. In contrast, image-based methods offer an accessible solution to construct panoramas that can be easily employed on a vast range of devices, including mobile phones. For instance, in the last few years many software tools such as PTGui [New01], Hugin [d'A07] or Microsoft Research Image Composite Editor [Mic12c] have been developed to compute automatic panorama stitching from images (see Figure 2.10 for an example). For this reason, in this research we

Figure 2.10: *Panorama stitched from 45 (*top*) and 56 (*bottom*) images using Hugin [d'A07].*

decided to employ image-based algorithms for constructing static panoramas, as such solutions can be easily replicated and used on a variety of hardware.

Image-based construction of panoramic imagery generally follows a two-step process. First, the arrangement of images to cover the panorama is discovered. Finding the arrangement of images is usually pairwise solved, by either direct or feature-based methods. Direct methods, such as the one proposed by Suen *et al.* [SLW07], search over the space of possible transformations between image coordinates to find the minimum pixel-to-pixel dissimilarities between the two images. Feature-based methods [BL03] use a sparse set of features to find correspondences between two images, from which they compute a transformation of image coordinates between the two views. Subsequently, images are combined to recover the final mosaic. The combination phase may include correcting for variations in lighting, color balance, and exposure. These techniques are readily available on smartphones. Diverdi *et al.* presented the Envisor system [DWH08] to construct a cube-map panorama by tracking SURF features, and Wagner *et al.* presented a system for constructing cylindrical panoramas by tracking FAST features [WMLS10].

Omnidirectional panoramas can be rendered in a variety of ways, with perspective and equirectangular projections being the most common solutions. Recent work [MSD+12] has explored the influence of varying projections on how users are able to locate scene objects. The work concludes that clear and understandable visualization of the panorama (i.e. equirectangular projection) is more important for whole scene object localization than maintaining real-world image features such as straight lines. In Chapter 5 implications of this finding will be discussed.

2.3.2 Spatio-temporal Media Exploration

Exploring large collections of unstructured images depicting the same location can be sometimes difficult or cumbersome. This process is often referred as "virtual tourism" or "surrogate travel" [Cla78], and its

Figure 2.11: Left*: the* Aspen Movie-map *experienced in the "Media Room" at the Architecture Machine Group, MIT, 1980. Photo credits Bob Mohl.* Right*: Lippman browsing the* Aspen Movie-map *on a touchscreen device. Photo credits Andrew Lippman [Lip80].*

early exploration dates as back as the late 1970s. A pioneer in this field was Andrew Lippman with his hypermedia system *Movie-maps* [Lip80]. At the end of the 70s, Lippman envisioned a system that would create, through videos and smart navigation, an experience of a remote location so immersive and realistic that newcomers would feel like they had already been there. Such system was originally commissioned by the DARPA's Cybernetics Technology Office, headed at the time by Craig Fields. The agency funded the project after Israeli soldiers practised for the recovery of an hijacked aeroplane by using an abandoned airfield made up to look similar. The training demanded a large preparation effort, which was mostly spent in recreating the remote airfield. Fields, then, requested a system that would facilitate soldiers training by creating virtual visit to new locations, so realistic and immersive that newcomers would literally feel as if they had been there before [Nai06]. Therefore, between 1978 and 1980, Lippman and Naimark recorded on videodisc hours of travels through Aspen, Colorado, with a camera mounted on the top of a car. The footage was then replayed with their system on large displays: users, also referred to as "traveller", watched the footage while seated in an instrumented armchair, controlling speed and direction of travel. Touch screens displaying map and aerial views allowed access to additional multimedia material, effectively enabling the first "surrogate travel" (see Figure 2.11). Later, several members of the original Lippman's team created movie-maps of additional locations, including the Paris Metro (1985), Palenque for the Bank Street College (1985), San Francisco for the Exploratorium (1987), Karlsruhe for the Center for Arts and Media (1990) and Banff for the Banff Centre for the Arts (1993).

During the last 35 years, Lippman and Naimark's seminal work have inspired a large body of research related to spatio-temporal media exploration, resulting in the development of many spatio-temporal photo visualization applications. *Photo Tourism* [SSS06, SGSS08] is one example, as the program aims to arrange and display a set of images in a 3D space so that spatially-confined locations can be interactively navigated. Similarly, the *PhotoScope* work of Wu *et al.* [WT09] extends the standard photo browsing paradigm by visualizing spatial coverage of construction site photos on a 2D map, and by indexing them with a combination of spatial coverage, time, and content specifications. Chen's *QuickTime VR* [Che95] consists of an image based system that uses 360-degree cylindrical panoramic images

to compose virtual environment and enable virtual spaces exploration. *RealityFlythrough* [McC07] uses videos combined with GPS and orientation data as its input. Videos are situated in a 3D representation of the world, allowing the user to navigate freely while continually transitioning to the most appropriate video for the current view. The system provides the user with some sense of how the videos relate to one another spatially, but no further context is provided and only one video is ever played back at the same time. To this extent, the 3D model of the scene is used as a map onto which videos are roughly located, and no further visual cues are extracted from it. McCurdy's system heavily relies on a property of the human visual system called "closure" [McC93], which is the brain's ability to fill in gaps when given incomplete information (in this case, the absence of visual information in-between views).

Unstructured video-based rendering [BBPP10] combines contemporaneous video streams of the same scene or performance, and provides an intuitive 3D-aware interface to these videos. It requires an image-based 3D reconstruction of the scene from photographs beforehand. This work was extended to try and model the dynamic foreground object as more than a billboard [TBP10], using volumetric approaches with moving-background-aware color models for segmentation. Tompkin *et al.* [TKKT12] introduce the *Videoscapes* system to explore sparse unstructured video collections. They build a graph of videos by visual similarity, exploiting this graph to generate 3D reconstructions at nodes, and then provide various different interfaces to explore this graph with seamless transitions. Dale *et al.* [DSAP12] introduce a system for browsing multiple videos with a common theme, such as the result of a search query on a video sharing website or videos of an event covered by multiple cameras. This browsing companion enhances a primary video by showing thumbnails of other temporally synchronized video clips.

Spatially-enabled exploration of single videos in isolation has also been researched. Hermans *et al.* [HVM+08] propose a visualisation for a single tripod video which presents most of the information of the original video in a single panoramic image. The authors capture a video on a pan-tilt head and then reconstruct a full static panorama from it. Dynamic foreground and background objects are segmented and decoupled in time to re-time motions in the original video footage. Finally, Pongnumkul *et al.* [PWC08] introduce a map-based storyboard system that presents a single tour video where the tour path is reconstructed, and coherent shots at different locations are pinned to a map.

Limitations

The works presented in this section assume the data to be organised in an "outside→in" structure: data where the cameras surround the subject of interest. Furthermore, they usually only show the spatio-temporal changes when transitioning between two videos at a time, and they require substantial additional data, such as photos, to reconstruct a geometric background model or a graph of hundreds of videos. Given their structure, these systems fail to achieve a good level of spatiality, conveying a very limited sense of space which is confined to the "action" in the scene. Clearly, these solutions are not suitable for our aims.

Without an enveloping context, in fact, spatio-temporal media exploration systems fail to show videos taken from the same place but with non-overlapping views of the scene, such as data from

Figure 2.12: *A Timescope and examples of different Berlin's areas seen through it. Image credits: Joachim Sauter (Permission to reproduce these figures has been granted by the original author).*

"inside→out" collections. This data structure, rather than focusing on a particular point or action, tends to obtain a wider description of an entire scene, conveying a higher sense of space while transmitting multiple actions within the same context. For these reasons, data available on-line, especially the geo-tagged data, is rarely organised using the outside→in paradigm, but rather the video collections are often structured in an inside→out manner.

Therefore, in our research we focus on this type of media organisation, as we believe it can encode higher spatial and temporal information of a remote location. However, the systems presented so far cannot structure, relate, and enable exploration of videos taken from the same place but with no visual overlaps, and therefore novel solutions are required. To overcome this limitation, in Chapter 6 we will introduce a system which is capable to spatio-temporally relate video-collections organised using the inside→out paradigm.

2.3.3 Focus+context Applications

Focus+context systems show a subset of information in full detail within a wider context of surrounding lower-density detail [CKB09]. An early exploration of this idea was performed by Naimark with his research idea *Time Binoculars* [Mic10]. With *Time Binoculars* Naimark envisioned binocular viewers that can provide enhanced information over a scene. In Naimark's prototype, small, high-resolution displays are integrated inside a binocular viewer, one for each eye, on axis with the binocular optics. The intensity of the display and binocular optics can be controlled, enabling the full range of transparency and opacity for both. Moreover, the unit's aiming mechanics, for panning, tilting, and zooming, incorporate sensors which can be used to determine the proper viewpoint for the displays. Small, high-resolution cameras may also be integrated inside the unit, one for each eye and on axis with the binocular optics. Finally, the unit may be connected via an on-board computer to the Internet. Thus, *Time Binoculars* enable on-site users to look around an actual site and see perfectly aligned augmentations of what they see, such as different times of day, different seasons or historical views. At the same time, a community of on-line users can watch and participate as well. While Naimark's idea never became a fully working system, a work similar in spirit was implemented in 1996 by the German non-profit organisation ART+COM [ART96]. The company installed around Berlin a number of binocular viewers, named *Timescopes*, which would allow people to experience where the Berlin Wall was located, how it divided the city and what it looked like (see Figure 2.12). Similarly to *Time Binoculars*, users could see through the units superimposed historical photos and films at their original location, enabling a virtual trip backwards in time.

The first research investigation in the focus+context domain has been performed for the first time by Baudisch *et al.* [BGBS02] with a system to enhance topographical map comprehension. A recent system, *CamBlend* by Norris *et al.* [NSQ12], extends focus+context interfaces to semi-panoramic (i.e. 180° FoV) video collaboration tools. A smaller focus window is moved around within a larger semi-panoramic video to identify objects to viewers of the scene. Contrary to the work presented in this thesis, CamBlend's presents only a single, small focus window, whose aim is to isolate individual objects rather than convey spatial information about the scene. When the focus window is employed, only the objects underlying the window are rendered in crisp graphics, while the rest of the scene is intentionally blurred. As such, the CamBlend video inset acts as a focussing window which isolates individual object rather than relating them together.

Neumann *et al.* [NYH+03] introduced "live" augmented virtual environments, where video from static surveillance cameras is projected onto geometric models from LIDAR data of a city. The goal of the system is to achieve coverage of the 3D model with video, and so have a walk-around 'live' virtual environment. Follow up work attempts live painting of the bare geometry environment with texture from video from a mobile observer [HYN05] and object extraction with background subtraction [SHYN03]. de Haan *et al.* [dHSdVP09] present a system for overlaying static security camera video feeds onto geometric models for virtual first-person viewing, similar to other later work [BBGP10].

Kim *et al.* [KOLE11] propose methods for augmenting aerial visualizations of Earth with dynamic information obtained from videos. However, the natures of the data (aerial looking down videos) dictate different and novel interaction tools. Additionally, the interactions in their work are speculative. In contrast, Chapter 6 introduces a user studies on interactions with videos in panoramic context, demonstrating significant improvements over existing video browsing techniques.

The methods introduced above have been extended to provide automatic camera control for tracking dynamic objects in virtual environments that have been augmented using multiple sparse static video feeds [SSM11]. Work similar in spirit by De Camp *et al.* [DSKR10] maps an indoor environment spatially top down, where each room is covered by one omnidirectional camera feed. Pirk *et al.* [PCD+12] enhance panoramas with embedded videos to create a new interactive medium. Videos are captured from tripods at the same time as the panorama is captured. Each video window occupies one region of the panorama and does not intersect any other videos. This work is distinct from the system presented in Chapter 6 as its goal is to enhance panoramas with dynamic objects, rather than to show the spatio-temporal relationships between videos by providing context.

Limitations

Investigating existing focus+context applications has been very important in inspiring this research as these applications form a superset for the videos in context problem. However, even if numerous focus+context applications exist in literature, none of them enables us to build the applications presented in Chapters 5 and 6, which form the backbone of this research. In Chapter 5 we introduce a teleconferencing system that uses smartphone cameras and panoramic imagery to create a surround representation of meeting places. In Chapter 6 we present a novel interface for spatio-temporal visualization and inter-

action within video-collection+contexts, using hand-held videos captured at different times.

The challenges set by these two applications, and consequently by our research, go well beyond what can be achieved with current focus+context applications. First, a crucial requirements for our applications lies in the interactivity of the solution. Especially for the teleconferencing system of Chapters 5, real-time alignment of videos to the context is crucial. As such, none of the presented solutions can fulfil this requirement. A second, important requirement is that the applications must deal with mobile video sources to obtain spatially coherent experiences. Again, none of the previous solutions can achieve this, limiting their application to static-camera scenarios. An additional requirement is that the systems need to obtain visually pleasant and geometrically corrected content rendering to avoid confusing the user. While some of the techniques presented here achieve this [PCD+12, NSQ12], the majority of the works present limitations in the rendering solutions that include visual artefacts, geometrically wrong texture re-projections or visualisations that do not map the original geometry of the scene. Finally, multiple video sources handling and rendering is also an important requirement, which however is not fulfilled by any of the previous works, limiting the conveyed temporal and spatial information.

2.4 3D Reconstruction

The vast majority of existing focus+context applications assumes that the scene geometry is know a priori in order to provide a valid context for spatio-temporal media exploration. This is a major limitation of such solutions, as acquiring the 3D structure of a scene is not always easy or even possible. 3D reconstruction is the process of capturing the shape and appearance of real objects, and it can be accomplished by either active or passive methods. Active methods physically interact with the object to reconstruct, either mechanically or radiometrically. Examples of active methods, such as 3D scanners or SL algorithms, have been already introduced in Section 2.2.2 and consequentially will not be covered here. Passive 3D reconstruction is usually performed with the aid of a (colour) sensor to measure the radiance reflected or emitted by the object surface and infer it shape. This section will cover such methods, which sometimes are also referred to as image-based reconstruction.

There is a vast range of solutions that tackle the reconstruction problem, and often when selecting a technique one has to sacrifice quality over speed or vice-versa. Broadly speaking, image-based 3D reconstruction techniques can be divided in three main categories: single-view reconstruction, multi-view reconstruction and structure from motion. However, almost none of such techniques can offer real-time depth acquisition as often the necessary post processing steps applied to the acquired data introduce a substantial computational overhead.

2.4.1 Single-View 3D Reconstruction

When viewing a common image, the human eye has no difficulty in understanding its 3D structure. However, this is a complex process combining both physiological (i.e. focus and accommodation) and psychological (i.e. Bayesian inference) cues, and therefore inferring the real structure of a scene from only one picture remains an extremely challenging task for current computer vision systems. The main difficulties are usually given by matching the local image features and their 3D location, an ambiguous

problem due to perspective projection.

During the last few years, there has been a lot of research focused on the problem of estimating 3D models from a single view. For example, shape from shading [ZTCS99, MWW02] and shape from texture [MR97, LG93, MP90] try to infer the 3D shape of an object by relying on purely photometric cues. These techniques proved to be fairly robust when applied to uniform textured objects, but they fail when the objects to reconstruct do not have uniform textures. Similarly to shape from shading, Criminisi *et al.* [CRZ00] presented a method to compute a 3D model by using solely geometric information determined from the image - a vanishing points with reference to a given plane.

Recently, Saxena *et al.* [SCN07, SSN07a] presented an algorithm to predict depth from monocular image features. Even though this work proved to be useful for tasks such as robot driving [MSN05] or improving performance of stereo-vision [SSN07a], it is not accurate enough to produce visually pleasing 3D reconstructed scenes. Delage at al. [DLN07, DLN06] and Efros and Herbert [HEH05a, HEH05b] presented similar works on reconstructing 3D shapes from one view, based on the assumption that the environment can be modelled with a flat ground and vertical walls. While the first authors focused on indoor images, the latter analysed outdoor scenes. They classified the image content into ground and vertical, producing a simple "pop-up" type fly-through from an image. However, even though these solutions produce visually-pleasant results, their reconstructed model are far from an accurate geometric reconstruction.

A different approach is introduced by Saxena *et al.* [SSN07b]. The authors present a work that focuses on inferring a detailed 3D structure that is both quantitatively accurate and visually pleasing. Other than local planarity, they make no explicit assumptions about the structure of the scene; this enables their approach to generalise well. Using a Markov random field they infer both the 3D location and orientation of the small planar regions in the image. They then learn the relationship between the image features and the location/orientation of the planes, and also the relationships between various parts of the image using supervised learning.

Finally *Dense Tracking and Mapping* (DTAM), a work by Newcombe *et al.* [ND10, NLD11], has recently gained a lot of interest in the the monocular stereo community for its promising results. DTAM is a system for real-time camera tracking and reconstruction which relies on "every-pixel matching" methods. As a single hand-held RGB camera flies over a static scene, the system estimates detailed textured depth maps. Hundreds of images are used to improve the quality of a simple photometric data term, and to minimise a global spatially regularised energy functional in a non-convex optimisation framework. The low computational time (DTAM achieves real-time performance using current commodity GPU hardware), together with robust camera tracking under rapid motion, make the system a hard competitor to the state of the art methods for single-view based 3D reconstruction. Recently, monocular dense 3D reconstruction has been ported from high-end computers [PRI+13] to portable devices [TKM+13].

Limitations

3D geometry has been leveraged in some videos in context works [NYH+03, HYN05, dHSdVP09] and it is typically employed in ICVE systems. However, such geometry requires a density and quality that

cannot be matched by the results typically available from the work presented in this section. Single view stereo offers poor reconstruction that cannot reconstruct the real scene in a suitable form for high quality telecommunication. Moreover, the geometry acquirable from one view is limited and often incomplete. An exception to this is DTAM [ND10, NLD11], which however relies on high-end computers and GPU hardware and builds un-texture models. Monocular dense 3D reconstruction running on portable and low powered devices is also possible [TKM+13], but in its current form this solution only allows reconstruction of small objects. Hence, more interesting solutions are offered by multi-view reconstruction.

2.4.2 Multi-View 3D Reconstruction

Multi-view 3D reconstruction, often referred to as multi-view stereo (MVS), is the process of reconstructing 3D geometry from a collection of images acquired from different vantage points. With the introduction of large, free image databases on internet (e.g., Flickr[Fli02] or Google Image[Goo01]), in the past years the research on MVS has seen a dramatic increase.

To organise and review the main MVS algorithms up to the year 2001, Dyer [Dye01] and Slabaugh *et al.* [SCMS01] have published two interesting surveys. However, due to the dramatic advances which are constantly made in computing, the state of the art in MVS keeps changing rapidly. In 2006 Seitz *et al.* [SCD+06] published an updated survey that reviewed and compare the latest MVS algorithms up to that year. Since then, other solutions have been presented, and therefore the best source of information on MVS remains the Middlebury benchmark [SS02], an on-line evaluation tool for MVS algorithms.

While there is a large body of prior work on multi-view stereo algorithms, we are mainly interested in those that, starting from a collection of images, are able to create and fuse depth maps and then build on top of these 3D models. To this extent, one notable work is the one presented by Narayanan *et al.* [NRK98]. The authors proposed the idea of creating dense shape models by volumetric merging of depth maps: after estimating individual depth maps through a traditional multi-baseline stereo matcher, they merge them to create a complete map that guides the final reconstruction. Starting from this work, Goesle *et al.* [GCS06] revisited the depth maps estimator to remove the high level of noise present in Narayanan *et al.* 's work: they developed a specialised matcher that computes depth only at high confidence points.

Following a different approach, Pollefeys *et al.* [PKVVG98, PVGV+04] use a three-step technique to recover 3D models from multiple views. They first perform a pair-wise disparity estimation for directly adjacent views which yields dense but incomplete depth maps. They then compute a joint estimate for each view by adding to partial maps corresponding disparity estimates from gradually farther away views on a per-pixel basis. Finally, the fused depth maps are combined by using a volumetric merging approach.

Lately, Furukawa and Ponce presented two important works on MVS. In [FP05], the authors present a method for acquiring high-quality solid models of complex 3D shapes from multiple calibrated photographs. After building a coarse surface approximation from the purely geometric constraints associated with the silhouettes found in each image, photo-consistency constraints are enforced in three consecutive steps. Firstly, places where the surfaces graze the visual hull are identified, then the visual hull is carved

using graph cuts and finally an iterative local refinement step is used to recover fine surface details. In their other work [FP07] they presented a dense multi-view stereo algorithm that incrementally builds a point cloud of the environment during the reconstruction process. From such reconstruction the authors build mesh representations using oriented points with normals by triangulating them using available algorithms such as Poisson surface reconstruction [KBH06]. The reconstruction is constrained by the quality of the reconstructed data point clouds, which are generally noisy and contain outliers difficult to remove from the final mesh.

A common assumption of many MVS works is that the images employed for the reconstructions are captured from viewpoints which are uniformly distributed across the scene. This allows the systems to perform a selection of k nearest images for each reference view, effectively allowing a global view selection that minimises occlusion and increases efficiency [AHES04]. However some works, such as [KKC+06], utilise more challenging and close to real-world data-sets in which the images are captured from non-uniformly distributed viewpoints. Such data-sets pose the problem of accurately selecting a subset of the available views with a uniform coverage of the space to reconstruct. Kang *et al.* [KSC01] solve this with local view selection, i.e. the assumption that temporal order of images matches spatial order.

Besides view selection, occlusion between viewpoints remains a challenging problem to solve. A number of recent stereo matching methods have used outlier rejection techniques to identify occlusions in the matching step [AHES04, NRK98]. Goesele *et al.* [GSC+07] further develop this approach and demonstrate that it can be generalised to handle many kinds of appearance variations beyond occlusions.

Another problem in MVS is posed by scene's lighting. A parallel thread of research in the stereo community is developing robust metric that can extend the view matching to several light conditions: variable [HK06], non-lambertian reflectance [JSY03] and with large changes in appearance [KKZ03]. Klaus *et al.* [KSK06] introduce a novel stereo matching algorithm that utilizes color segmentation on the reference image and a self-adapting matching score that maximizes the number of reliable correspondences. The scene structure is modelled by a set of planar surface patches, and a disparity plane is assigned to each of them. The optimal disparity plane labelling is then approximated by applying belief propagation [Jud82], yielding to accurate results. A work similar in spirit is presented by Wang and Zheng [WZ08]. The authors introduce a new stereo matching algorithm based on inter-regional cooperative optimization. The proposed algorithm uses regions as matching primitives, defining the corresponding region energy functional for matching by utilizing a) the color statistics of each region and b) a constraints on smoothness and occlusion between adjacent regions. The experimental results on the Middlebury test-set [SS02] indicate that the performance of these last two region-based methods are competitive with most state of the art stereo matching algorithms.

Finally, work of Liu *et al.* [LCDX09] showed how estimating depth maps in a continuous manner, in contrast with the vast majority of work that use a discrete model, yields very good and detailed results. Their work, starting from depth estimation done by integrating silhouette information and epipolar constraint in a variational, continuous model, generates depth candidates that are first refined on a path-based

normalised cross correlation (NCC) metric, and then merged in a final global map.

Limitations

MVS algorithms offer solutions that, in term of acquired geometry, are more suitable to our work than the ones offered by single-view stereo. Still, the reconstructed models are often quite poor in details. Higher-quality models can be achieved, albeit a dramatic increase in the computational effort required. While this is not a problem for scenarios that can afford off-line processing and heavy computation effort, it becomes a major limitation for our work, which aims to leverage portable devices and a variety of different and technically asymmetric configuration to recreated remote destinations.

In addition we note that the vast majority of MVS algorithms require careful setup, which includes multiple camera arrays or specific hardware, long and tedious calibration process, and controlled lighting and environmental conditions. This is perhaps a stronger hindrance for MVS, which limits its application to very specific scenarios, and goes against the BEAMING's minimal technical intervention principle. Therefore, MVS is largely unsuitable for the work presented in this thesis.

2.4.3 Structure from Motion

Single-view and MVS reconstruction are usually constrained to work with one or multiple view that come from a set of static, calibrated cameras. When this constraint cannot be satisfied, a valid solution to perform scene reconstruction is given by Structure from Motion (SfM). SfM is a CV process that aims to simultaneously reconstruct the unknown 3D scene and camera positions and orientation from a set of images acquired from a moving camera. Typically, in a SfM system the input is an image sequence from a moving camera and the output is a 3D geometry describing the underlying scene and camera motion. As the camera is uncalibrated and no scene information are given, SfM is typically a much harder problem to solve than single-view or MVS reconstruction.

Finding structure from motion presents a similar problem than MVS reconstruction. In both cases, the correspondence between images and the reconstruction of 3D objects needs to be found. As such, SfM systems are heavily dependent on image features correspondence (please refer to [MTS+05, TM08, Sze06] for detailed reviews of the most diffuse view-invariant image descriptors and their matching algorithms). To find correspondence between images, features such as corner points [Mor83] or SIFT descriptors [Low04] need to be tracked from one image to the next. The feature trajectories over time are then used to reconstruct their 3D positions and the camera motion. Having information on camera positions and motion and image points correspondence means that a point cloud describing the underlying scene can be reconstructed through point triangulation.

While seminal work based on only two frames for SfM has been presented by While Lounguet-Higgins [LH81] at the beginning of the 1980s, the multi-frame SfM techniques still widely used today occurred more than a decade later. These methods include the global optimisation techniques by Tomasi and Kanade [TK92] and the factorisation algorithm introduced by Spetsakis and Aloimonos [SA91], Szeliski and Kang [SK93] and Oliensis [Oli99].

More recently, related techniques to photogrammetry (i.e. bundle adjustment) have been brought to CV, and they are now regarded as the best solution to perform 3D reconstruction from multi-view point

correspondences. This is the case of the work by Triggs *et al.* [TMHF00] and Hartley and Zisserman [HZ04]. However, while these two approaches rely on using algebraic techniques, the work of Szeliski and Kang [SK93] can handle perspective projection and partial or uncertain tracks by using a non-linear least squares, sparse matrix based technique which quickly converges to the desired solution.

The previously mentioned techniques work under the assumption that the camera calibration parameters are known; when this is not true, self-calibration techniques such as the ones presented by Pollefeys *et al.* [PKVG98] and Pollefeys and Van Gool [PVG02] can be used. Self-calibration is the process to estimate a projective reconstruction of the 3D scene and to then perform a metric upgrade of it.

The SfM approaches described so far were not designed to deal with large and heterogeneous data sets. The first work to handle large, heterogeneous dataset has been the one of Brown and Lowe [BL05], but only with the work of Snavely *et al.* [SSS08, SGSS08, SSS06] and Agarwal *et al.* [AFS+11], unordered and completely random selected images have been employed. In both cases, Internet image databases have been used to retrieve images featuring the same subject captured under different viewpoints and lighting conditions. Based on these researches, Microsoft has recently developed *Photosynth* [Mic08], a software that recreates fully navigable, image-based rendered 3D scenes from user inputs.

Due to the heavy computation required by most SfM solutions, ways to accelerate the process have been recently investigated by the research community. In particular Frahm *et al.* [FFGG+10] expanded and optimised the work in [AFS+11] by improving the performances and the computation time of their algorithm. Even if using a single machine, the authors are able to reconstruct a dense 3D scene from over 3 millions pictures in a single day. They force the algorithm to respect geometric and appearance constraints to obtain a highly parallel implementation on modern graphics processors that allows for fast computational times without sacrificing the quality of the reconstruction.

When dealing with non-rigid scenes, the works presented in this section handle the reconstruction by assuming various constraints about the scene (i.e. non-rigid shape bases, scene deformation, the shape itself or about the camera motion). However, these additional constraints limit the practical applicability of the methods. A solution to this limitation has been introduced by Dai *et al.* [DLH12] by proposing a novel and simple solution to non-rigid factorization. The proposed method does not assume any extra prior knowledge about the problem other than the low-rank constraint, hence it is "prior-free". Nevertheless, it does not suffer from the basis ambiguity difculty, but is able to recover both camera motion and non-rigid shape accurately and reliably.

Limitations

Similarly to MVS, SfM techniques require intensive computational power and can be sometimes too slow to be used directly in real-time telecommunication. In addition, typical output of SfM is a sparse point-cloud which can hardly be employed to describe remote environments if is not augmented with additional visual information. While this is possible, we note that similar results can be achieved in real-time with consumer depth-cameras (see Section 2.6.2), and therefore in our research we decided to employ the latter solution. However, SfM offers a good understanding of the scene, especially in terms of camera poses and calibration, and therefore, if off-line processing can be afforded, it presents a valid

solution to calibrate different cameras and perform more accurate multi-camera MVS reconstruction.

2.5 Content Rendering

Rendering quality has been found to have a significant influence on presence and task performance in VE systems [Zim04, ZP03], ans as such, rendering quality of both 2D and 3D graphics is a crucial task in any ICVE or VMC system. For this reason, content rendering plays an important role in the research presented in this thesis, and therefore during my studies I have explored several works on this subject. Given the type of data involved in the research (video streams and point-based representations), two particular rendering techniques are mostly suited to our work: image-based rendering (IBR) and point-based rendering (PBR). Both fields have been extensively explored in the last few years, and especially IBR, due to the growing availability of high-definition cameras and on-line video databases, is becoming again a topic of great interest in the CG and CV communities.

In the rest of this section we will present the most relevant works on IBR. During the research we also performed an investigation of the PBR literature, and experimented with several existing solutions. However, we established that given the complexity of such solutions, the computational costs associated with them grows exponentially when applied to large models. Therefore, given the dense sampling offered by the hardware employed throughout this research and the real-time constrains set by our applications, we decided to opt for a less sophisticated point description and rendering technique, based on classic point-clouds and OpenGL rendering instructions, which however allows for real-time handling and rendering. The readers interested in this topic can refer to Kobbelt and Botsch detailed survey on point-based representations [KB04], and to Gross and Pfister book "Point-Based Graphics" [GP07].

2.5.1 Image Based Rendering

IBR is a rendering technique that aims to reproduce 3D-like environments with the aid of a set of images, seamlessly fused together on top of an existing 3D structure (often referred to as "proxy geometry"). IBR, firstly introduced by Chen and Williams [CW93], has been developed for high quality realistic representations of static scenes [LH96, GGSC96], but recently it has been extended to directly employ video footage [BBPP10] in less constrained environments [DLD12, BBM+01].

During the years, many solutions have been proposed to achieve photo-realistic IBR. However, given the nature of the research topic here presented, the focus of this section can be restricted to two categories: *Light field* based and *lumigraph* based rendering techniques. The main difference between these two approaches is in the sampling and rendering of the *plenoptic* function - a function of five variables representing the flow of light at all positions in all directions. By densely sampling such function, light fields and lumigraphs can provide a faithful reproduction of 3D scenes.

Typically, light field or lumigraph based systems implement a 3-step pipeline that includes an acquisition stage, a camera pose estimation algorithm and a rendering phase. For the literature related to the second step please refer to Section 2.4.3 and 2.6.2. A number of approaches have addressed the acquisition stage, including robotic arms [LH96], camera arrays [WJV+05] or microlens array for single-camera light fields ([AW92, NLB+05, GL10]). A capture approach closer to casual, real-world data acquisition

has been introduced by Gortler et *et al.* 's for their lumigraph system [GGSC96]. The authors propose a specially-designed stage for pose estimation of a hand held camera. The user though must gauge the density required for good coverage, which is usually difficult to achieve for hand-held capture of light fields and lumigraphs. Buehler *et al.* [BBM+01] ad Davis *et al.* [DLD12] try to overcome this limitation during their rendering and acquisition steps, respectively. Both solutions offer visual information that help the user to densely cover the reconstruction space. For instance, Davis *et al.* provide a visualisation technique that, during acquisition stage, interactively informs the user on which viewpoints will be later available for reconstruction.

Once acquired, the plenoptic function is used to synthesise novel viewpoints of the scene. This process is sometimes also referred to as view dependent texture mapping (VDTM). The basic approach to VDTM is put forth by Debevec *et al.* [DTM96] in their image-based modelling and rendering system called *Facade*. *Facade* is a system used to create geometric models that resembles a set of input images. As part of this system, a rendering algorithm was developed where pixels from all relevant cameras were combined and weighted to determine a view-dependent texture for the derived geometric models. A real-time, hardware-accelerated version of this algorithm has been proposed by Debevec *et al.* in [DYB98]. Other solutions for VDTM have been proposed for the lumigraph [BBM+01] and light field [LH96] rendering systems, but due to their complex algorithms these solutions cannot work at interactive rates.

The rendering algorithm used for unstructured lumigraphs [BBM+01] selects views based on a variety of criteria, such as angular distance, and then uses a k-nearest-neighbour reconstruction that ensures that the interpolation weights for each pixel fall off smoothly to zero for its k^{th} nearest neighbour. This approach comes with some limitations. For instance, scalability can hardly be achieved, since for each sample on the image plane the penalties for every input image must be evaluated and sorted to find the k-nearest neighbours. Another problem is that the k-nearest neighbours might exhibit a poor angular distribution around a given location (e.g., the nearest neighbours may not surround the reconstructed view or they may all be on one side and then suddenly switch to the other side as the virtual camera is moved). Furthermore, the blending field may have discontinuities and is not always monotonic as a function of the distance to a viewpoint projection because of the normalization term. A different rendering approach is proposed by Lipski *et al.* [LLBM09], based on triangulation of the input cameras. However, the authors use simple bilinear interpolation over the entire output image, restricting the sampling of reconstructed views to linear combinations of just a few input views. Davis *et al.* [DLD12] overcome this limitation by integrating Lipski *et al.* 's technique with the one adopted for the unstructured lumigraph system. In this work the authors present a new rendering algorithm that is tailored to the unstructured yet dense data captured by the user. Such method can achieve piecewise-bicubic reconstruction using a triangulation of the captured viewpoints and subdivision rules applied to reconstruction weights.

Finally, also the work of Snavely *et al.* [SGSS08, SSS06] provide a broad but sparse coverage of the plenoptic function by combining large photo collections. The strength of their approach is the ability to leverage photos that have already been taken. They can, for example, acquire a miniature object by rotating it in front of a camera. However, their approach is not suitable for interactive applications.

Limitations

Most of the works proposed in this section focus on static images rendering, and therefore are not directly applicable to real-time scenarios. Additionally, typical data employed in IBR are usually captured from a homogeneous and static set of cameras, further limiting the possible scenarios. However, some solutions allows for casually captured data to be employed, such as unstructured video-based rendering [BBPP10], lumigraph [BBM+01] and lightfield [BBM+01], albeit prohibitive computational time and hence, non-interactive rates.

Clearly, a direct application of the methods described in this section is not suitable for both this research and the BEAMING platform. Nevertheless, regarding the latter, the principles described in [BBM+01] and [DLD12] have been a valuable source of inspiration for the rendering techniques described in Section 4.2.

2.6 Depth Fusion

With the term depth fusion we indicate the task of merging depth measurements of a scene, as seen from a monocular camera, into a global depth-map. This is somehow similar to the well studied problem of Simultaneous Localization and Mapping (SLAM) task. Even though depth fusion has been investigated for quite some time, is only with the introduction of low priced depth sensors that fusing and improving depth samples has becoming a popular topic of research.

Depth fusion is highly relevant to the research presented in this thesis, as the reconstruction techniques developed during the research rely on monocular depth acquisition from range sensors. The rest of this section will present an overview on the most common depth fusion techniques. It will first introduce works related to depth fusion for sensor improvement, and it will then present the most relevant RGB-plus-depth (RGBD) systems.

2.6.1 Depth Fusion for Depth Sensors Improvement

Due to the limited field of view and working range of the majority of the available depth sensors, depth fusion is a crucial task when reconstructing large models with range cameras. Feulner *et al.* [FPKH09] propose a simple solution to consecutive ToF camera frames registration by detecting edge presence in the intensity image and aligning their 3D coordinates by maximising the non-centred correlation coefficients. In contrast, Fuchs and May [FM08] filter points at depth discontinuities that have the largest distance error, while Swadzaba *et al.* [SLP+07] present a full acquisition pipeline that improves depth accuracy with the use of several preprocessing steps such as distance-adaptive median filter applied to the intensity, amplitude and depth image to remove points with low amplitude and a neighbourhood consistency filter that detects and removes noisy pixels at edges location.

By fusing high resolution colour images with ToF depth and colour frames, Yang *et al.* [YYDN07] highly enhance depth-maps and fuse them in a single, global map. In their work, to reduce the level of up-sampled blur that occurs at discontinuities areas, a bi-later filter is applied to aggregate the probabilities of estimating correct depth based on colour segmentation (i.e. pixels located in the same colour segment should have similar depth estimate).

Expanding the *LidarBoost* work of Schuon *et al.* [STDT09], Cui *et al.* [CSC+10] achieved state-of-the-art 3D reconstruction results. By randomly selecting a ToF camera generated point cloud, other point clouds, and thus depth maps, can be aligned to it. This is achieved by centring at each point a multi-variate Gaussian, and by estimating maximum likelihood through Expectation Maximization. Another relevant work for real-time depth fusion has been presented by Merrell at al. [MAW+07]. Their method selects depth estimates for each pixel that minimises violations of a visibility constraints and thus removes errors from the depth maps. A two-stage process is performed to fuse several depth maps: the first stage generates potentially noisy, overlapping maps from a set of calibrated images and the second stage fuses these depth maps to obtain an integrated surface with higher accuracy and minimal noise and redundancy.

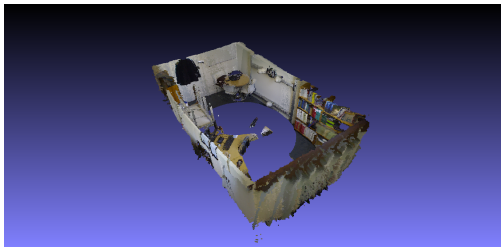
Finally, Reynolds *et al.* [RDP+11] have recently proposed a technique to improve ToF depth samples, by employing a per-pixel confidence measure built using a random forest regressor trained with real-world data. The authors claim that their heuristic improves over previously developed metrics that use the amplitude of each ToF sample as a measure of confidence. This is supported by results based on two different ToF sensors, showing how an improved confidence measure leads to superior reconstructions in subsequent steps of traditional scan processing pipelines.

Given the dense data acquired by SL cameras, as well as the rich colour information available, the methods presented in this section can be discarded in favour of image-based and feature-based mapping techniques. Such techniques, faster and sometimes more robust than the one mentioned above, are introduced in the next section.

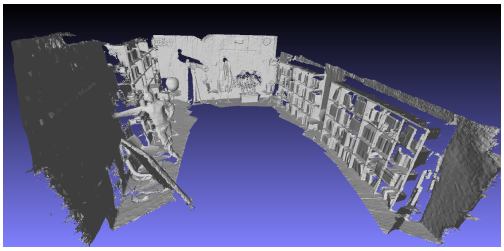
2.6.2 Depth Fusion for Environment Mapping

Due to its fast performances and appealing results, environmental depth mapping, also know as depth fusion, has been chosen as the main 3D reconstruction technique for the BEAMING static model acquisition. Reconstructing large environments using depth fusion techniques allows us to fulfil several requirements of the BEAMING platform. Besides its interactive rates, another advantage introduced by depth fusion is the employment of point-based data structures to hold scene information. Point-based representation of the scene allows fast manipulation for dynamic content changes. In fact, as we explore tracking and reconstruction in the context of user interaction, it is critical that the representation we use can deal with dynamically changing scenes, where users directly interact in front of the camera - a non-trivial requirement for a 3D reconstruction system. For instance, previous work [MAW+07, ND10, NLD11] employ mesh-based representations for live reconstruction from passive RGB sensors, but unfortunately they do not deal with changing dynamic scenes.

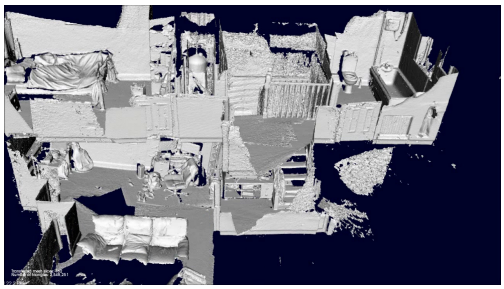
A solution to this problem is given by an environmental depth mapping technique called “RGBD-mapping”. RGBD systems are able to reconstruct dynamic scenes at an interactive rate by employing a point-based, or mesh-based, representation of the scene that can dynamically change. RGBD systems [HKH+10, HJS08, HSJS10, EEH+11] rely on continuous and robust detection of sparse scene features to align monocular depth-maps into a global map of the scene. Given the nature of the data used (colour plus depth), the global depth-map can be easily described with a sparse points collection, obtaining inter-



(a) RGBDSLAM *[EEH+11].*



(b) KinectFusion *[IKH+11]. Please notice the high quality mesh, albeit no colour textures and loop closure.*



(c) Kintinuous *[WKF+12] (Permission to reproduce this figure has been granted by the original author).*

(d) Kintinuous *extended with surface colouring [WJK+13].*

Figure 2.13: *Results of three RGBD mappers.*

active rendering rates, efficient memory management and data manipulation. Typically, a RGBD system exploits both colour and depth information to register consecutive frames into a global map. The visual features extracted from the colour images are used to roughly register consecutive frames in 2D, while the depth information is employed to extend the obtained transformation to a 3D coordinate system. The resulting transformation is used as a starting point to robustly align consecutive frames. By either employing an Iterative Closest Point scheme (ICP - see [RL01] for a study) or Procrustes analysis on the input data, the two frames can be quickly merged. Based on these principles, Huhle *et al.* [HJS08] present a scene acquisition system which allows for fast and simple acquisition of arbitrarily large 3D environments using a range camera. In each step of the processing pipeline, colour and depth data are used in combination to gain from different strengths of the sensors. A novel registration method is introduced that combines geometry and colour information for enhanced robustness and precision. Henry *et al.* [HKH+10] introduce *RGB-D Mapping*, a full 3D mapping system that utilizes a joint optimization algorithm combining visual features and shape-based alignment. The authors exploits visual and depth information for view-based loop closure detection, followed by pose optimization to achieve globally consistent maps. However, this additional step reduces the reconstruction frame-rate to ∼2Hz. Also Engelhard *et al.* [EEH+11] introduce an environmental mapping system, named *RGBDSLAM*, that acquires large environments at interactive frame-rates (Figure 2.13(a)). Unlike other systems, the authors apply an additional optimisation step after the ICP alignment to optimize the acquired pose graph, using a pose graph solver.

RGBD mapping is a powerful tool to fast reconstruct large and complicated environment. However, as it heavily relies on visual features, its application is limited to feature-full environment. To overcome

this limitation, a different approach in terms of alignment algorithm and scene description is introduced by Izadi *et al.* [IKH+11]. The authors, with their system termed *KinectFusion*, present a framework that, similarly to other RGBD systems, creates detailed reconstruction of indoor scenes in real time (Figure 2.13(b)). The novelty of *KinectFusion* is that the systems uses only the depth data from a Kinect camera, and therefore no explicit visual feature detection is needed to build a global map. Moreover, the system does not build a dense depth-map, but instead reconstructs a "growing" surface which more accurately approximates the real-world geometry. *KinectFusion* reconstructs 3D models in real-time, and to do so it uses a heavily GPU-powered pipeline. Such pipeline consists of 5 stages: depth-map acquisition and conversion to real world space, camera tracking through a GPU-based ICP step, volumetric integration to update the running surface (i.e. a voxel grid) and volume ray-casting to extract the view to render to the user. *KinectFusion* has proven to be extremely efficient and accurate in reconstructing large environment, becoming in short time the state of the art solution for the depth fusion task. However, unlikely [HKH+10, EEH+11], KinectFusion does not tackle the "loop closure" problem when dealing with reconstruction of closed environment (e.g., rooms).

Loop closing is the task of deciding whether or not a sensor has, after an excursion of arbitrary length, returned to a previously visited area [NH05]. Since drifting in the camera location estimation is almost impossible to avoid, closing a loop often results in re-optimising a set of camera locations such that the first estimate will match the last one. Reliable loop closing is both essential and hard, and it is without doubt one of the greatest challenge for long term, robust RGBD mapping. However, such problem is not novel, as it has been largely studied in the Simultaneous Localization and Mapping (SLAM) literature (see [Ho07] for a detailed study). The hard part about loop closing is not only asserting the presence of a loop, but also detecting when loop closure is even a possibility. To solve this, visual features are often used, as they proved to offer the best performances in terms of accurately measuring the amount of appearance change between camera views [ZLY10]. Newman and Ho [NH05] employ a mixture of visual features, temporal information and scanning laser data to estimate the probability that a loop is imminent and needs to be closed. Ho and Newman [HN06] extend this solution by employing image features that are both visually salient and wide-baseline stable. To achieve fast loop closure, the authors build an image-based retrieval system where each frame is time stamped, processed and finally entered into a database as a collection of feature descriptors which are optimised to achieve fast comparison. To further fasten this process, Callmer *et al.* [CGNR08] employ "tree of words", a delayed state information filter and planar laser scans for relative pose estimation. The authors claim to achieve loop closure in near real time, with a false detection rate of about 0.01%.

When a loop is detected, the whole pose graph must be optimised. Henry *et al.* strategy for loop closure is to represent constraints between frames with a graph structure, with edges between frames corresponding to the relative transformations given by the initial alignment step [HKH+10]. To keep their graph relatively sparse, they introduce the concept of "key-frame", determining them on visual overlaps of frames. After they align a frame F, they reuse the image features to find a rigid transformation with the most recent key-frame, using the same RANSAC [FB81] procedure defined for the frame-to-

frame alignment. As long as the number of RANSAC inliers is above a threshold, F is not added as a key-frame to the pose graph. Every time a new key-frame is added, the system checks for potential loops, and when one is detected, the whole graph is optimised. Engelhard *et al.* [EEH+11, EHE+12] fasten this optimisation technique by applying it to a graph that has been already optimised several times during the acquisition step. Every time a new frame is added, a local optimisation of the graph is performed, allowing for the final pose optimisation to converge extremely fast.

Building on these works, *KinectFusion*'s limitations have been partially solved by Whelan *et al.* with their system called *Kintinuous* [WKF+12, WKLM13] (Figure 2.13(c)). This is achieved through a) altering the original *KinectFusion* algorithm such that the region of space being mapped can vary dynamically, b) extracting a dense point cloud from the regions that leave the volume due to this variation and c) incrementally adding the resulting points to a triangular mesh representation of the environment. The authors' approach incorporates a number of enhancements over the original *KinectFusion* framework, including the integration of multiple 6-DoF camera odometry estimation methods for robust tracking and loop closure. In a folow-up work, Whelan *et al.* extended the *Kintinuous* framework to perform real-time surface colouring [WJK+13] (Figure 2.13(d)).

Limitations

KinectFusion is indeed a notable example of depth mapper. However, it does not provide coloured and closed meshes due to limitations of its algorithm. *Kintinuous* tries to solve these limitations, producing higher quality results. However, the output of such framework grows in size exponentially, and transmitting such models over the Internet may pose a big challenge. On the contrary, the RGB-mapper systems presented in [HKH+10, EEH+11] provide coloured and geometrically-closed point cloud, are both reliable and fast, and employ a point-based representation which can be easily compressed and transmitted. Hence, as the main scene reconstruction tool used for the BEAMING platform, we decided to employ *RGBDSLAM*, modifying the original implementation to add extended functionalities (see Chapter 4.2).

2.7 Chapter Summary

This chapter has been divided into six main sections. The first section presented an overview of visual telecommunication systems, establishing the importance of video-mediated communication over verbal-only communication by describing state-of-art VMC systems and positioning them as the optimal form of high-quality interpersonal remote interaction. VMC's inherent problems with regards to representation of 3D space, along with novel but imperfect approaches aiming to alleviate this problem were then detailed. ICVEs were then presented as a maturing medium, able to overcome the spatial limitations of VMC, locating users in a navigable and interactive shared graphical environment populated with objects and avatars embodying users. The topics of immersion, spatiality and presence, central to VE systems, were also covered.

The second section focused on video acquisition and transmission. Panoramic and depth-enabled solutions, along with their limitations, were detailed and discussed. The work was presented also with respect to the design choices made for the BEAMING platform, and especially for depth-cameras, their

suitability for fast acquisition and transmission was discussed. To this extent, the topic of depth streaming was also covered.

The third section illustrated the most relevant work on static panoramic imagery acquisition and introduced work related to spatio-temporal media exploration and video+context applications. Panorama stitching techniques were introduced, followed by a discussion on spatio-temporal exploration of large video collection that included historical work, as well as state of the art solutions. The section concluded with an introduction to focus-and-video+context applications, with an illustration of pioneering ideas in the topic and more recent and high-quality solutions to the problem.

The fourth section focused on the main techniques developed for 3D reconstruction for large environments. Single-and-multi-view stereo techniques were discussed, alongside with their limitations. MVS' inherent problems with regards to speed and accuracy alongside with novel approaches aiming to alleviate these problems were then detailed. The well established research on structure from motion was also covered, illustrating how this technique can complement and improve MVS reconstruction.

The fifth section discussed aspects related to content rendering, with a special focus on image-based rendering. The main IBR techniques, together with their limitations with regards to speed and suitability for ICVEs, were illustrated.

The sixth and final section explored work related to data fusion for large environment mapping, presented here as a valid alternative to standard SfM and MVS reconstruction. Depth fusion for sensor improvement and environment mapping was introduced, with a description of the most important RGBD mappers whose development was highly boosted by the recent introduction of inexpensive commercial depth cameras. Common problems related to this type of system (i.e. loop closure and mesh consistency) were introduced alongside with the solutions proposed by the research community.

The following methodology chapter is divided into two sections. Firstly BEAMING, the ICVE platform which was (collaboratively) developed and used in some of the following experimental work, is presented. Secondly, the specific camera hardware employed during my research are introduced. In particular, negative and positive aspects of each solution are described. To support the discussion, the chapter ends with a qualitative comparison analysis of three depth-camera technologies.

Chapter 3

BEAMING: An Asymmetric Telepresence System

In theory, there is no difference between theory and practice. But in practice, there is.

Manfred Eigen

This chapter introduces the reader to BEAMING (Being in Augmented Multimodal Naturally Networked Gatherings [Con10]), the main project that motivated the research presented in this thesis and under which a large part of the development was done, and to the specific hardware used during it. These concepts are fundamental to understand the work presented in the rest of the thesis and the forthcoming experimental work.

Section 3.1 introduces the reader to BEAMING, which acts as the primary ICVE platform supporting some of the experimental work investigated over the following chapters. The main platform's ideas, as well as more detailed description of various components and hardware employed in the system, are presented and discussed. A particular focus on the high-level technical aspects of the systems is introduced. However, an in-depth description of the two specific platform instances which have been developed and tested during the research are presented in Chapter 4 rather than here.

Section 3.2 introduces the reader to the specific camera hardware employed in BEAMING, and consequently in this research. The hardware include an omnidirectional camera and three depth-enabled cameras. Finally, a comparison of the three depth cameras is presented, with a focus on their technical and qualitative aspects. Please note that some of the images used in this chapter are adapted from the author's own work [SSO+12].

3.1 The BEAMING System

Most collaboration tools such as VMC systems and CVE platforms provide symmetric access to a shared medium. For example, in videoconferencing, each person usually sees a view of the other participants and their surroundings. Although these systems can be configured similarly to face-to-face meetings, they lack some of those meetings' immediacy. As already noted in Section 2.1, researchers have argued that this is mostly due to the systems' technical limitations, presenting ICVEs as an alternative that

supports full 3D shared spaces and that consequently can better mimic real face-to-face meetings.

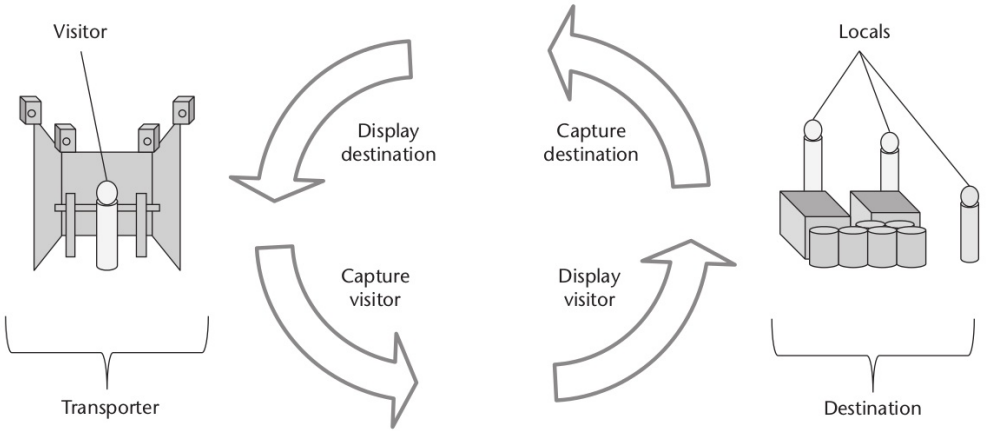
Nevertheless, ICVE technologies tend to be laboratory based and are still far from being widely distributed. Therefore, participants normally cannot access these systems without leaving their work places or living spaces. Additionally, such technologies generally feature technological symmetry to ensure that the same sensory cues are available to all parties. Symmetric communication systems require each of the system's end to be equipped with similar, if not identical, hardware solutions, in order to guarantee similar access to a shared content and user experience. The BEAMING project tackles technological and access issues head on. The platform abandons the symmetry of access to a shared virtual environment in which collaboration happens, and rather focuses on recreating, virtually, a real environment and having remote participants visit that virtual model. To this extent, the display systems can be in any reasonable space such as an office or meeting room, domestic environment, or social space, and can support any level of fidelity and immersion. Therefore, BEAMING represents an asymmetric communication system in which different ends of the system can be equipped with varying hardware solution, which can greatly differ in terms of quality and complexity. At the same time though, BEAMING grants similar social affordance and sensory cues to all connected users, regardless of their technical setup.

3.1.1 System Overview

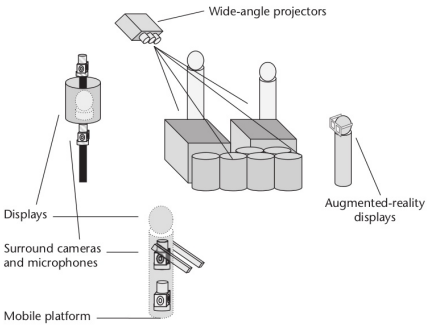
To better explain the BEAMING concept, I will introduce the reader to a typical platform's application. Imagine a lecture to be given by a professor physically located several miles away from the university's lecture theatre. Currently, the only options for the professor is to either physically travel to the place or to have a video-mediated conference. BEAMING replaces these solutions by allowing the professor to "beam" into the lecture theatre from his home or office and give the lecture to the students through his virtual representation at the destination. Fans of the popular science fiction television series *Star Trek* will find this concept familiar. The professor can be represented via a physical robot or a virtual avatar and viewed by the students on dedicated displays or through AR-based visualisations. Interaction between the professor and the students will still be possible via their relative virtual representations. Moreover, through the BEAMING replay facility, other students can become passive spectators days after the lecture has taken place.

Figure 3.1(a) gives an high-level overview of the system. In a typical BEAMING session, the "destination" is a real space where people, called the "locals", interact. The transporter is a high-end VR system equipped with 3D surround visuals, 3D surround audio, tactile and haptics systems, and biosensing. The transporter's user is here defined as the "visitor", a person located in a different physical location than the destination. BEAMING's capture and display strategies are bi-directional, as the system aims to capture the destination and display it to the visitor and simultaneously capture the visitor and display him or her to the locals.

As already discussed before, one goal of BEAMING is that the destination should not be a laboratory space with carefully calibrated equipment, but rather any physical space where hardware can be introduced. As such, any technical intervention must be portable or mobile, self-calibrating, and dynamically configurable. It should also be as unobtrusive as possible so that it does not interfere with the



(a) *BEAMING recreates a real environment (the destination) populated with people (locals). A remote participant (the visitor) visits that virtual model via the transporter.*



(b) *Types of technical intervention at the destination.*

(c) *A professor "beams" in a another laboratory to meet a colleague miles away form his office.*

Figure 3.1: Top*: The BEAMING system overview.* Bottom - Left*: A possible BEAMING's technical setup at the destination.* Bottom - Right*: An artistic depiction of a typical BEAMING application: virtual meeting. Image credits [SSO+12].*

locals' behaviour. The "destination-visitor" paradigm in BEAMING is fundamentally technologically asymmetric but aims to support symmetric social interaction between the visitor and locals. One way of achieving this is by exploiting the objects at the destination, as these are key mediums through which the social interaction takes place. An example of a potential technical interventions is given in Figure 3.1(b). Here, the destination is equipped with mobile robots, situated displays, wall or environment projections, camera capture and audio capture. At the same time, a locals wear augmented reality glasses to enrich their visual experience.

With respect to a typical BEAMING session, the scope of the research documented in this thesis covers all tasks that are concerned with creating and transferring a multi-sensory, visual experience of the destination to the visitor. Such tasks include visual capture, representation, transmission and rendering of the destination environment.

3.1.2 An Asymmetric System for a Symmetric User Experience

The major goal for the BEAMING system is to provide a collaborative mixed-reality environment that grants symmetrical social affordance and sensory cues to all connected users whether they are locals

(a) *A visitor wearing a HMD and a motion-tracking suit.*

(b) *A local at the destination, seeing visualizations of the visitor on a spherical display and a wall projection. A Kinect camera tracks the local, and a surround camera is next to the sphere.*

Figure 3.2: *BEAMING's mediating technologies are highly asymmetric between the destination and transporter sites.*

or visitors. In other words, BEAMING aims to be an asymmetric system that supports symmetric user experience (see Figure 3.2). Although the mediating technologies are highly asymmetric between the destination and transporter sites, visitors' behaviour should not be hindered because of their remote location. Also, they should be represented to the locals with a virtual or physical embodiment. Borrowing terminology from the VE field, we may say that BEAMING strives to give a sense of spatial presence within the destination for visitors and a sense of co-presence among both locals and visitors. With respect to the work presented in this thesis, the research focuses on the first aspect, and investigates ways to enhance visitors' spatial presence at the destination with various forms of scene acquisition and rendering.

An important role in the BEAMING platform is given to the visual display at the visitor site. Such displays ideally must be immersive, such as a head-mounted display (HMD) or a display similar to a CAVE (Cave Automatic Virtual Environment [CNSD93]), but also other types of display are investigated and employed in this research and in the BEAMING platform. This is because, to strive for social symmetry, the system should provide similar sensory experiences, particularly the dominant visual mode, to all parties. As locals, perceiving the actual physical location, need no visual mediation to perceive the destination as being realistic and spatial, stimuli representing the destination must be transmitted in real time to the visitor site. These stimuli need to depict the destination as accurate as possible, as they are essential to ensure that the dominant visual mode is similar to both parties. This first challenge is partially covered by some of the work developed in this research.

However, besides visual acquisition, also the technological display properties used by the visitor are in important factor to ensure a similar sensory experience, As such, they must foster the impression of being physically at the destination. While this is an important factor to reach a symmetric user experience, and one would expect that immersive presentations increase the social symmetry in an asymmetric and heterogeneous system architecture, in this research we also investigate spatial representations and rendering solutions that can convey an adequate sense of space when rendered on non-immersive,

as well as fully-immersive displays. In the experimental work presented in Chapter 7 implications on users' spatial reasoning of using different display modes coupled with surrounding representations are investigated and discussed with a controlled user study.

3.1.3 System Requirements and Hardware

Achieving a social symmetry in an asymmetric, heterogeneous system architecture sets many technical challenges. In the BEAMING project, a variety of researchers are tackling, in a largely collaborative effort, many of these challenges in a range of modalities including robotic telepresence, visual, audio and haptics representations, novel display types and emotion recognition and display.

The project's most ambitious goal is perhaps the will to reconstruct the whole destination in real time. To do so, a variety of specific technologies are required, and solutions to integrate and interact with them have been proposed during the project's life. In the rest of this section, examples of specific hardware and technical demonstrators built during the BEAMING projects are reported. A more specific description of how these have been integrated into two BEAMING platform instances will be given in Chapter 4. As the research presented in this thesis, with respect to BEAMING, mainly focuses on the visual acquisition of the destination, a detailed description of the camera hardware employed in BEAMING is given in Section 3.2. Specifically, the research presented here contributed to the BEAMING project by investigating and developing a variety of solutions to acquire, reconstruct and stream visual descriptions of the destination and the locals. This not only enabled the candidate to investigate challenging technical issues and develop novel algorithms, but it also allowed him to conduct a scientific evaluation on how people perceive space through videos in panoramic context while interacting with remote users, and how well they can understand and act upon the representation.

Networking

BEAMING aims to facilitate long distance communications, and therefore the link between visitors and locals relies on data transmission between sites. Linking remote sites across public switched networks such as the Internet is particularly prone to delay, jitter, disorder and loss of packets. However, responsiveness to viewpoint updates is a key requirement for immersive graphics that, if insufficient, may reduce the sense of presence [MRWB03] and result in motion sickness [CNSD93]. A way to increase the responsiveness of viewpoint updates is through a loosely-coupled replicated database approach, albeit reduced synchronisation of observed events, especially when they originate from distinct sites. However, responsiveness of objects and avatars during interactions does not need to be as high, even though low response levels can impact on the interpretation of non-verbal communication, and are likely to hinder both conversational and object-focused interaction.

For these reasons, the BEAMING's networking layer, mainly developed by Oyekoya Oyewole and William Steptoe at UCL, adopts a client-server replicated shared object database (henceforth known as the BEAMING Scene Server - BSS), based on discrete shared objects described by numeric or string-based properties. Data transmission is managed by RakNet middleware [Ocu10], a cross-platform C++ game network engine that provides UDP transport. A detailed description and evaluation of the server is given in Oyekoya *et al.* [OSS+13].

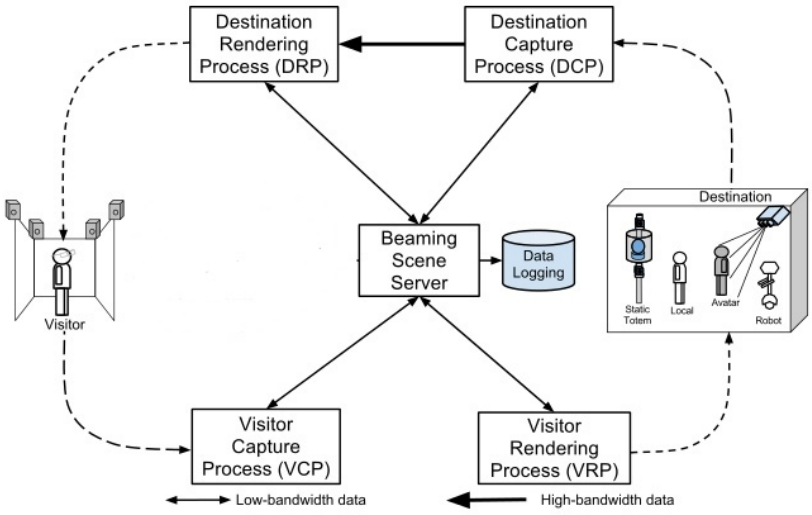


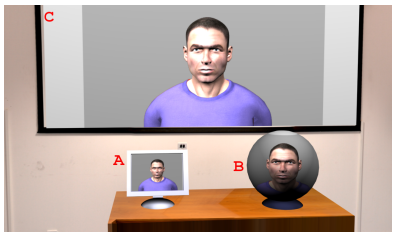
Figure 3.3: *BEAMING Scene Service configuration for a typical BEAMING session. Image [OSS+13] (Permission to reproduce this figure has been granted by the paper's authors).*

The BSS does not explicitly handle video streaming, which is one of the contributions of this research (see Chapter 4). However, the server has information related to the acquisition devices present in the scene, and holds calibration information for the entire camera network. For completeness, please note that the video streaming network is based on VP8 compression [Goo11] and has a client-server structure which, similarly to the BSS, is handled by RakNet middleware.

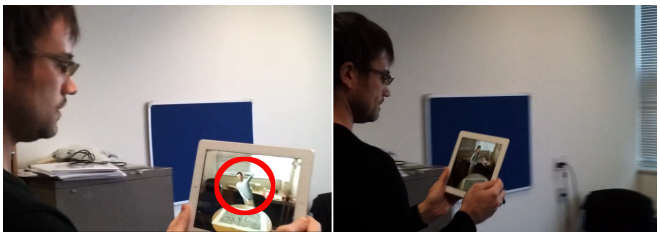
Clients connect to the BSS if they wish to access the shared object database. Once connected to the BSS, each client process will receive a copy of the shared objects. Clients may also create objects and publish updates, which are then replicated across all connected clients. Objects may be queried, retrieving any updates since the last query (this is typically performed in the client application loop). Each object is associated with a particular data type. There are two classes of data: a) tracked human, typically represented by an avatar or robot, holding information on skeleton position, facial expression and tactile feedbacks and emotions; and b) reconstructed environment, typically represented by video, audio, point cloud or objects, holding information about audio and video sources, point cloud location and 3D objects.

Most of the data that are exchanged via the BSS arise from visitor capture processes. The primary purpose of this data transfer is to facilitate the visualisation of the visitor at the destination site. However, rendering processes might receive data from several sources. The data flow of the capture process occurs at a high rate and low size (i.e. motion tracking, physiological streams). Slow rate and low size data are also broadcast, which includes configuration information and status information of devices (e.g., their current sending rates or detection of features in the signals). At the destination, multiple rendering processes are required due to the multiple output modes, including visual, aural, and haptics modalities. These are the data that are sent from the visitor to the destination, and are typically lesser than the data sent via the other path.

Figure 3.3 describes the BSS configuration for a typical BEAMING session. As a visitor client is



(a) *Display types - A) SCREEN-avatar; B) ICON-avatar; C) PROJECTION-avatar.*



(b) *Display types - AR-avatar: in red, the virtual avatar embedded in the tablet's back-facing camera view.*

Figure 3.4: *Visualisation of destination display types.*

represented as an avatar at the destination, its process creates as many shared objects as there are avatar joints. This client will own all of these objects, and will perform local manipulation on the properties. In this case capture processes (Visitor Capture Process in the figure), such as motion capture collocated at the client, will update positions and rotations. These changes are then serialised to keep all client databases consistent. Subsequently, a client display located at the destination site (Visitor Rendering Process in the figure) will receive new updates and animate the avatar accordingly.

To enable replay of BEAMING sessions, log files are recorded. This logging is performed on the server, and writes all node updates to a human-readable file that is time-stamped from a central time server (Data Logging activity in Figure 3.3). In addition, the server also records an audio file of all participants' talk in OGG-Vorbis format [Xip00]. The logging process is essential to enable later playback, and log files may be replayed on both non-immersive and immersive displays.

Displays

Displays technologies in BEAMING are an important factor to reach a symmetric user experience. As the platform's mediating technologies are highly asymmetric between the destination and the transporter, the variety of display types greatly differs between sites. Thus, displays can be grouped in destination displays and transporter displays.

Representation of the visitor is a major challenge for BEAMING, and a mixed-reality approach is proposed for the destination displays, which include four technologies, with the generic term "avatar" used to denote the representation of the visitor in the destination (see Figure 3.4). The available representation modalities are:

- **SCREEN-avatar**, which simply uses local screens to display an avatar of the visitor and is the most basic form of visitor display. display size are not limited, and can range from a laptop display up to a large flat screen.

- **ICON-avatar (or Sphere-avatar)**, which is a spherical display (Global Illumination's Magic Planet [Glo06]) that allows rotation of a displayed avatar head to face any point in the destination. Orientation information are obtained at the transporter and sent to the client driving the display. A detailed description and evaluation of the Sphere-avatar is given in Oyekoya *et al.* [OSS12].

- **PROJECTION-avatar**, which are projections on to surfaces in the destination, featuring a large

projector located at the destination site. The size of the display allows for greater visibility of body language and a wider range of physical movement.

- **AR-avatar**, which is an AR portable device display, which embeds a virtual avatar representing the visitor into the destination live-view acquired from the back-facing device camera. The client driving the display uses camera tracking based on image features, allowing the system to run in marker-less environments.

Regarding the locals, they can be represented at the visitor's side using pure 3D graphics, by embodying the local in an avatar representation, using surrounding video mapped to a sphere or employing a hybrid approach featuring embedded 2.5D video of the locals within a VE. As these solutions have been largely investigated by this research, a detailed description of them is given in the following chapter. Capturing of the locals is designed to operate within immersive and non-immersive displays, including:

- **CAVE**-based systems;
- **HMDs**, such as the NVis nVisor SX111 [NVI08] or the Oculus Rift [Ocu12].
- Large or medium-sized **flat displays**.

Tracking Devices

The BEAMING platform currently supports a number of tracking devices which are used for capture of elements of a visitor's activity during a BEAMING session. However, if available, tracking information may be captured also at the destination site. These devices can be grouped in users' position and orientation trackers and limb trackers and emotion recognition systems.

Position and orientation trackers include:

- **Kinect skeletal tracker**: a skeletal tracker based on the range data acquired by a Microsoft Kinect camera. Two distinct solutions are supported, one based on the Kinect for Windows SDK [Mic12a], and one based on the OpenNi SDK [App10].
- **OptiTrack and Optitrack V120 Trio** (NaturalPoint) [Nat96]: an optical motion capture systems from NaturalPoint that enables real-time and high-fidelity near full-body skeletal motion capture of an individual, excluding finger and toe joints.
- **Fastrak** (Polhemus) [Pol00]: a magnetic tracking system integrated in the platform through the Virtual Reality Peripheral Network (VRPN) library.
- **IS900** (InterSense) [Int96]: an acoustic tracking systems that provides extremely fast user tracking.

Emotions recognition systems and limbs trackers include:

- **Glove** (Essential Reality) [Ess02]: a low-cost and accurate finger and hand tracking system.
- **Viewpoint Eye Tracker** (Arrington Research) [Arr95]: an eye tracker that uses infra-red to illuminate the eye and track the position of the pupil from a video stream.

- **Enobio** (Starlab) [Sta11]: a wearable, modular and wireless electro-physiology sensor system for the recording of EEG (Electroencephalogram - brain activity), ECG (Electrocardiogram - heart activity) and EOG (Electrooculogram - eye movement).

- **Faceshift** (Faceshift AG) [Fac12]: a face performance capture system for capturing face motions. The motions, captured with a depth-camera, are described as a mixture of basic expressions, plus head orientation and gaze and are then used to animate virtual characters.

Tracking is mainly developed and investigated by UCL, University of Barcelona (UB) and Starlab. The Kinect-based tracking solution is the system's preferred choice of tracking, given its affordability and compactness. Such solution, contrary to the ones based on dedicated tracking systems, requires only a single camera and low computational power to perform a high-rate and precise multi-user tracking. This solution, then, is in-line with BEAMING's minimal impact technical intervention goal.

Audio

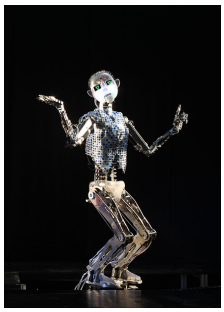
Verbal communication between visitors and locals is a critical aspect of BEAMING. To reach an immersive audio experience, the platform supports 3D audio effects (also knows as binaural audio), a group of sound effects that manipulate the sound produced by stereo speakers, surround-sound speakers, speaker-arrays, or headphones. 3D audio involves the virtual placement of sound sources anywhere in three dimensional space, which in BEAMING are attached to each local or visitor.

We will now outline the locals' speech capturing and rendering. However, as audio communication is symmetric, the principles outlined to playback locals' speech at the visitor site are similar to playback of visitor's speech at the destination site. Audio capturing, mainly developed by the Acoustics group of Aalborg University, is performed by a head-mounted microphone given to each local. At the destination a computer is equipped with a multi-channel sound card (RME Hammerfall DIGI 96-8PST with an optical interface to a Behringer ADA8000) to capture the audio from these microphones, which is done by a BSS client using PortAudio [Por06]. The audio from head-mounted microphones is sampled using the 8-channel sound card, compressed using audio codecs such as Opus [ICWG12] and is then transmitted through the local network to be picked up by the audio server. Tracked position and rotation of both the visitor and the locals are required to achieve believable binaural audio. As tracking information of the locals is already published on the BSS, each local is given an unique ID which is attached to each particular microphone stream. To avoid latency, only single channel sound is passed through the network. By processing a single channel using a head-related transfer function (HRTF), such channel can be turned into two channels, which then serve as input to the listener. The visitor client has access to the required tracking information to correctly place audio sources in the visitor's virtual 3D space.

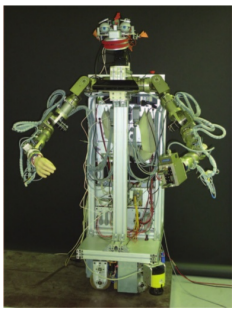
Robots and Haptics

Besides dedicated displays, in a typical BEAMING's setup a visitor can also be physically embodied with a robot avatar. Robotic aspects of the platform are mainly investigated by the Institute of Automatic Control Engineering of the Technical University of Munich (TUM). However, preliminary investigations are carried out also by UCL and UB. The robots employed in BEAMING include:

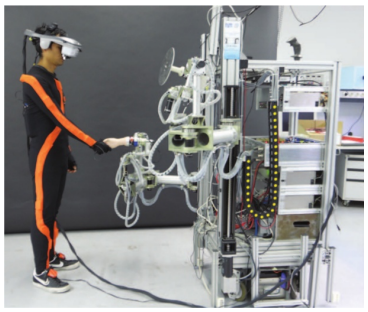
(a) *The QB Anybots. Image credits [Any02].*



(b) *The RoboThespian. Image credits Engineered Arts Limited - press kit.*



(c) *A robot avatar that mimics the visitors movements and emotions. Image [SSO+12].*



(d) *A visitor interacting with an encountered-type haptics device that mimics the form of the destination and the locals' interactions. Image [SSO+12].*

Figure 3.5: *Examples of robotic and haptic devices used in BEAMING.*

- **QB Anybots** (Anybots, Inc.) [Any02] (Figure 3.5(a)): a mobile robot featuring a small display for video-chat interaction and a dynamically balancing platform that enables wide-ranging mobility.
- **RoboThespian** (Engineered Arts Limited) [Lim07] (Figure 3.5(b)): a life sized humanoid robot designed for human interaction in a public environment. It is fully interactive and user-friendly, but its movements are limited to upper limbs (i.e. the robot cannot move around a room).
- **Mobile robot avatar** (TUM) (Figure 3.5(c)): an anthropomorphic robot with two robotic arms and hands and an emotion-expressing head. The visitor's arm and hand movements are tracked by a motion capture suit and mapped to the robot's movements. The system driving the robot also analyses the visitor's facial expression and recognize emotional states, which the emotion-expressing head then conveys.

Besides robotic telepresence, BEAMING also aims to integrate haptics feedbacks and rendering. Haptics investigation is mainly conducted by the Perceptual Robotics Laboratory (PERCRO) of the Scuola Superiore Sant'Anna, Pisa and TUM. Haptics devices include:

- **Haptics Vest** (UB): a haptics vest developed at UB, comprising of a velcro vest, an array of vibrators (up to 25) and a micro-controller board, that provides haptics feedbacks to the visitor.
- **Encountered-type haptics devices** (TUM) (Figure 3.5(d)): haptics devices that let the visitor perceive interaction with the objects at the destination or with locals.
- **Finger-mounted portable device**: haptics device that can display the transition between contact and non-contact of the fingers.

3.2 Cameras

One of the most ambitious goal of BEAMING is to reconstruct the destination in real time. To do so, we decided to employ video-based and depth camera-based solution to build a network of heterogeneous devices which are able to capture the destination in real time. For this reason, an important part of this

research focuses on analysing and manipulating different types of camera. While a detailed description of the approaches taken towards this task is given in Chapter 4, the rest of this chapter will describe the camera hardware involved in this research. Moreover, Section 3.2.5 will present a comparison of three depth cameras technologies in terms of qualitative and quantitative results to motivate some of the design decision made during the development of the various BEAMING platform instances.

3.2.1 PointGrey Ladybug3

(a) *Camera hardware.*

(b) *A panoramic texture as acquired by the Point Grey Research Ladybug3.*

Figure 3.6: *The PointGrey Ladybug3 camera.*

The Ladybug3 camera (Figure 3.6(a)), developed by the Canadian company Point Grey Research [Poi10b], is a camera capable of generating live omnidirectional videos. This camera combines the views acquired by six 2 megapixel (MP) Sony CCD sensors into a single, panoramic view which results in having a 12 MP resolution. In contrast to other cameras, the Ladybug3 directly streams to disk the raw, mosaicked images acquired by the six different sensors. Therefore, the process of de-mosaicing, converting and stitching together the different views is entirely done on software on the host machine. A more in-depth description of the unit's stitching technique is presented in Section 2.2.1. This solution, while giving much more control on the acquired data, creates a large computational overhead during surround video acquisition. The device, when working at full resolution, is capable of recording surrounding video with a frame rate of 15 frames-per-second (fps). This results in a potential bandwidth requirement of 90 MB/sec ca. (assuming that each pixel is described with 8-bit). In terms of contribution to the BEAMING platform, the camera provides an easy way to capture a surrounding video of the destination (see Figure 3.6(b) for an example of such texture rendered using an equirectangular projection.). When this is coupled with immersive displays the visitor, as will be clear in the following chapter, can experience an immersive representation of the destination. Immersive projection technologies such as CAVEs or HMDs feature a wide field-of-view, allowing for people to use natural movements to look around a remote location. The full 360° images provided by the Ladybug3 then, when combined with such display types, provide a more natural way of exploring the scene than any other conventional 2D video.

In Chapter 5, the impact of using the Ladybug3 for teleconferencing and remote collaboration is explored. Results from the study show that the omnidirectional video can highly enhance users' sense of

(a) *Camera hardware.*

(b) *Bumblebee XB3 depth-map. Colour-map varies from black (farther) to white (closer or invalid).*

Figure 3.7: *The PointGrey Bumblebee XB3 camera.*

spatiality, providing a higher level of spatial information which users can easily understand and act upon. Therefore, the omnidirectional video offered by a Ladybug3 is a compelling visualisation that can easily replace a more geometric accurate 3D reconstruction of an environment and that can be easily captured, visualised and streamed across networks.

3.2.2 PointGrey Bumblebee XB3

The Point Grey Research Bumblebee XB3 [Poi10a] (Figure 3.7(a)) is a depth-enabled camera that allows the acquisition of stereo pairs, rectified view reference and disparity map of a scene. The camera's hardware consists of three sensors that can reach the maximum resolution of 1280×960 pixels, working at 15 fps. The unit employs a stereo matching algorithm to estimate depth information of a scene (see Section 2.2 for more details). As for the Ladybug3, the camera requires additional software to perform rectification, disparity estimation and de-mosaicking of the original stereo pairs, but it does not require any calibration process. An example of the data produced by the camera can be seen in Figure 3.7(b). The employment of the Bumblebee XB3 can strongly benefit the description of a scene. Besides depth-map, the camera offers large coloured textures and can be potentially used for tracking users. Point Grey Research also offers a software solution to calibrate a network of multiple units.

Similarly to the Ladybug case, the amount of data produced by the Bumblebee camera is considerably large and can reach, when considering the RGB rectified image obtained at full frame-rate, a throughput of 50MB/sec ca. . When combined with the surround video, the Bumblebee can allow the placement of extra content in the scene such as 3D objects or avatar representations of remote users. This allows for more realistic user experience of the depicted scenes. As an additional feature, a rough reconstruction of the underlying scene can be obtained through the available depth-map, which can be easily converted into a point-based representation (i.e. a point cloud). However, being entirely based on stereo reconstruction, the Bumblebee struggles to evaluate depth values in textured-less areas (see Figure 3.7(b) for a depth-map example). Additionally, the camera requires specific settings for each scenario, making its usage cumbersome and limited. Unfortunately, these are a substantial impediments which drastically limit the employment of the camera for robust and reliable depth acquisition.

3.2.3 Microsoft Kinect & ASUS Xtion PRO Live

The Microsoft Kinect technology is based on the classic SL approach. The unit comprises of two cameras, one RGB and one IR, and one laser-based IR projector (Figure 3.8). The IR camera and the IR

Figure 3.8: *Sensor placement within a Kinect sensor. The baseline is of approximately 7.5cm.*

projector form a stereo pair with a baseline of approximately 7.5 cm. The IR projector sends out a fixed pattern of light and dark speckles. The pattern is generated from a set of diffraction gratings that are designed to lessen the effect of the zero-order propagation, i.e. to avoid a centered bright dot [ZSMG07].

As already described in Section 2.2.1, depth calculation is performed by triangulating the known pattern emitted by the projector, that is stored on the unit. For each new frame, depth is estimated at each pixel p_i by sliding a correlation window on the recorded IR frame. The window is typically small (9×7 or 9×9 pixels). It is used to compare the recorded pattern at p_i with the corresponding stored pattern. The best match gives an offset from the known depth, in terms of pixels, also known as disparity. The device performs an interpolation of the best match to get sub-pixel accuracy to $\frac{1}{8}$ of a pixel. Given the known depth of the stored pattern and the disparity value, an estimated depth for each pixel is calculated by triangulation.

Since the camera requires constant projection of the infra-red pattern into the scene, combining multiple Kinect cameras is non-trivial due to potential interference problems. To combat this, one could carefully align multiple units to avoid IR overlaps, but this would require a tedious manual calibration. A more general solution, based on constant shake of the units, has been recently presented by Butler *et al.* [BIH+12]. The authors propose to associate to each unit a motor with an offset weight. The motor shakes the Kinect, and subsequently the shacking also moves the IR projector and the IR camera. As the shaking is constant for both the IR projector and sensor, the depth estimation algorithm still works reliably for the single unit. However, from the view-point of another Kinect, the dot pattern of the shaken projector moves around and interferes with its own pattern only for a small amount of time. This results in a reduced interference between cameras.

A different solution to mitigate interference errors is introduced by [BRB+11]. The authors apply a set of fast rotating disks to multiple Kinect units, effectively creating a time division multiple access (TDMA). Each disk contains a gap large enough to allow a laser beam to pass through it. Hence, each unit's laser diode is blocked by the disk, except for the time when the gap is allowing the laser to project its pattern into the scene. Each Kinect ™is equipped with such a disk rotating at the same speed but with a different phase, ensuring that only one laser projects the patter into the scene at any given time.

Similarly to most range cameras, the Kinect suffers from systematic error in depth estimation.

(a) *A Kinect unit.* **(b)** *A Xtion PRO camera.* **(c)** *Kinect depth-map. Colour-map varies from black (farther or invalid) to white (closer).*

Figure 3.9: *The Microsoft Kinect and ASUS Xtion PRO cameras.*

Interestingly, the error seems to be stronger when depth measurements are collected near the camera sensor [SJP11]. There are several approaches to handle the systematic error, including the one presented in [SJP11]. Herrera *et al.* [HKH12] propose a distortion model to correct the systematic unit error. A different approach is introduced by [YHMY12]. The general principle beyond their calibration routine is that, as the SL principle is based on both emitter and receiver, the intrinsic parameters of both the IR camera and projector should be taken into account. Hence the authors present a depth correction model that is based on joint estimation of depth-camera and projector intrinsic parameters, achieved by only showing a planar board to the depth sensor.

An example of the data produced by the camera can be seen in Figure 3.9(c). The hardware consists of two cameras that output two videos at a frame rate of 30 fps, with the RGB video stream at 8-bits VGA resolution (640×480 pixel) and the monochrome video stream used for depth sensing at 16-bits VGA resolution with 2048 levels of sensitivity. The camera working range is between 1.2–5.0 meters. However, realistically the reliable working range of the camera is between 1.2–3.0 meters, as the random error of depth measurements increases quadratically with increasing distance from the sensor, and reaches 4 cm at the maximum range of 5.0 meters [KE12]. The unit also contains a multi-array microphone that allows speech recognition. The raw data (i.e. colour-plus-depth) produced by a single Kinect can reach a bandwidth of 25MB/sec when run at maximum resolution and full frame rate. The camera also features a motor that enables vertical tilting.

Due to the popularity of the Kinect camera, ASUS and PrimeSense decided to release a similar device, named ASUS Xtion PRO Live (see Figure 3.9(b)). The two companies released a range of depth-enabled cameras, with the ASUS Xtion PRO Live being the closer to a Kinect device. The differences between the two cameras hardware is that the Xtion lacks the microphone array, and therefore has a smaller body size, and has no tilting motor, hence it is entirely powered through USB. Table 3.1 illustrates the qualitative differences between the two cameras. Due to its compact size, better portability, lower weight and better RGB image, during the research we have been often opted for the ASUS Xtion PRO camera. However, it is important to note that the drivers and API used to interface with the Xtion are the same that are used to pilot a Kinect, and therefore software written for one camera works reliably also with the other device.

Both Kinect and Xtion cameras can offer similar information to the one available from a Bumblebee XB3. However, the precision of depth estimation and resolution of images is inferior to the one offered by the Point Grey camera. While the latter limitation can be solved by fusing several depth-maps together,

	Pro	Cons
Kinect	• Stable work with various hardware models • Has motor that can be controlled remotely • Has array of microphones	• Bigger size (12" x 3" x 2.5") • Higher weight (1.0 Kg) • Require ACDC power supply • Higher interference with another Kinect sensor • Worse RGB image quality
Xtion PRO Live	• More compact (7" x 2" x 1.5") • Lighter weight (226 g) • Does not require power supply except USB • Lower interference with another ASUS Xtion Pro • Better RGB image quality	• Does not work with some USB controllers (especially USB 3.0) • No motor, only manual positioning

Table 3.1: *Qualitative analysis of the Microsoft Kinect and ASUS Xtion PRO Live sensors.*

the quality issue is highly dependent on the application. Based on our experiments and applications however, we can conclude that a good depth approximation can be achieved using SL cameras, with quality matching the requirements of the scene reconstruction needed in BEAMING.

3.2.4 PMD[vision] CamCube

The PMD[vision] CamCube (Figure 3.10(a)) is a time-of-flight camera that features a PMD chip with a resolution of 200×200 pixels. This means that the camera works with more than 41 thousand distance values for each frame at a rate of up to 25 frames per second, generating, for the depth stream only, a bandwidth of 4 MB/sec ca. (depth values, unlike the Kinect and Xtion case, are described with a 32-bit notation). The camera operates at a standard modulation frequency of 20 MHz, which results in an unambiguous range of about 7.5 meters. As any other camera system, the PMD[vision] CamCube camera can suffer from over-saturation in case of too long exposure times in relation to the ambient background light and the objects' distance and/or reflectivity. However, the integrated suppression of background illumination (SBI) provides the CamCube with an enhanced dynamic range so that it can operate even in bright environments. While the camera is able to produce depth values densely and with high precision, it only captures grey-scale intensity images (a colour coded version of the range image), making the integration of colour images extracted from other cameras necessary if a textured 3D model is required. An example of the data available from the camera is illustrated in Figure 3.10(b).

ToF cameras are active imaging systems that use standard optics to focus the reflected light onto the chip area. Therefore, the typical optical effects like shifted optical centers and lateral distortion need to be corrected for, which can be done using classical intrinsic camera calibration techniques. This applies also to the CamCube camera. However, as the camera has a low resolution which is rather small in comparison to standard RGB- or grayscale-cameras, standard calibration techniques have to be applied

(a) *Camera hardware.*

(b) *Depth-maps available from a PMD camera. Colour-map varies from black (farther or invalid) to white (closer).*

Figure 3.10: *The PMD[vision] CamCube camera.*

with care [LK06]. Similar to the Kinect case, the parallel use of several cameras may lead to interference problems, i.e. the active illumination of one camera influences the result of another camera. This kind of interference can be circumvented by using different modulation frequencies.

The CamCube, while offering a limited resolution depth-map, is certainly the unit that offers the most precise and reliable depth measurements. The depth samples obtained from a CamCube, in fact, correspond to the real world depth values, as the camera exploits per-pixel light information, unlike other technologies that sample the space in a regular grid and then up-sample the obtained map to match their frame resolution. However, as already mentioned before, the quality of the depth-map is highly dependent on the application. To this respect, we believe that the CamCube, on its own, is not usable for direct reconstruction, due to its low resolution and lack of colour data.

3.2.5 Depth Cameras Comparison

To conclude the camera hardware overview, in this section I will introduce a comparison of the three depth cameras presented above, in terms of their qualitative properties and their suitability in VR/AR systems. Positive and negative aspects of each camera will be analysed together with a final discussion on what is the solution adopted in BEAMING.

Analysis

Table 3.2 summarises the analysis presented in this section. The Bumblebee camera has a large frame size (1280×960 pixels) and a maximum frame-rate of 15 fps, supported by a large reliable working depth range (0.5–4.5 meters encoded with 16-bits or 32-bits depth pixel). Moreover, the three RGB sensors are pre-calibrated, as the underlying technology is based on stereo algorithms. However, the camera requires careful parameters tuning for each new environment and struggles in retrieving depth information in non-textured regions. The entire depth calculation is performed on the host machine, and the unit is also expensive. The Kinect and Xtion Pro cameras' positive aspects are manifold. They are low-priced, with a good reliable working range (1.2–3.0 meters [KE12], described with 16-bits depth pixel), an average resolution (640×480 pixels) and a maximum frame-rate of 30 fps. Most importantly, the camera provide depth measurements under a large variety of light conditions. The depth values are directly available from the hardware. Even though the depth frames acquired from the cameras need to be registered with the RGB frames (the two sensors are located apart from each other), this task is

computationally inexpensive and is supported by the driver's API. The PMD[Vision] CamCube acquires depth information directly on the camera hardware. Virtually no calibration is required, but it can be used to improve depth estimates. This camera has several limitations: besides being expensive, the PMD unit can only acquire grey-scale images with a very limited frame size (200×200 pixels) and a maximum frame-rate of 25 fps. Moreover, due to the technology employed, the depth measurements can be very noisy and affected by the ambient light, limiting its use to indoor scenarios. However, it is important to note that this is also the case of most SL cameras, and especially for the Kinect case, errors in the measurements are mainly related to the lighting condition, which influences the correlation and measurement of disparities [KE12].

	Pro	**Cons**
Bumblebee	• Large frame size • Large working range (0.5–4.5 m) and very large depth range (16 or 32 bits precision pixel) • No calibration or frame registration required	• Expensive • Depth values computed on the machine • For each scenario it needs an ad-hoc setting • Problems in retrieving depth in non-textured regions
Kinect/Xtion	• Inexpensive • Depth values directly from hardware • Works under large variety of conditions • Good working range (1.2–3.0 m) and large depth range (16-bits depth pixel)	• Depth frame needs to be registered to the RGB frame • Average frame size (only VGA) • Use of multiple cameras can be difficult due to IR sensors interference
PMD	• Large working range (0.5–7.5 m) and very large depth range (32-bits precision pixel) • No calibration or frame registration required • Depth values directly from hardware	• Currently still expensive • Only gray-scale image • Very limited frame size

Table 3.2: *Qualitative analysis of depth cameras.*

Conclusion

In the near future, ToF cameras will not only be extended to support colours and higher frame sizes, but also rapidly drop in price, as confirmed by the recent release of the second version of the Microsoft Kinect, which is based on ToF technology. Moreover recent work [BBK07, KS06], prove that ToF cameras generate more accurate depth estimation than any stereo vision solution, especially in highly dynamic environments, such as a typical BEAMING session. For the time being, however, inexpensive structured-light cameras are an attractive off-the-shelf solution to perform depth estimation with limited noise and good accuracy, and in fact they are the current devices that are used at the core of BEAMING's 3D reconstruction technique. Due to their unique properties, stereo cameras will remain a valuable addition to any camera network, being able to augment the scene reconstruction with precise depth

measurements in region where SL and ToF cameras may not provide sufficient details, such as areas with occluding edges [KS06].

3.3 Chapter Summary

This chapter introduced the reader to some concepts fundamental to understand the research presented in this thesis. The chapter firstly introduced the main BEAMING idea, an ICVE system that aims to enable telepresence in a variety of modalities and with a vast range of hardware. BEAMING's main goal is to provide an asymmetric system for a symmetric user experience. In other words, BEAMING's users will be able to perceive the same experience, no matter if real or virtual, with hardware setup that can considerably vary across sites.

The chapter continued with an introduction of the hardware employed in BEAMING and a detailed description of the networking solution employed in the platform, being this perhaps one of the most crucial aspects of an effective ICVE system. Audio, robotics, haptics and display devices have been described to give to reader a sense of the hardware used in the platform instances that will be introduced and discussed in the following chapter.

Moving into specific work related to this thesis, the chapter continued with a detailed description of the camera hardware employed in the platform. This section is decoupled from the hardware description presented in the first part of the chapter as during the research, which has focused on the reconstruction of the destination, we investigated and evaluated solutions to exploit the cameras presented here to reconstruct in real-time the destination at the visitor site. To motivate the camera choices, we have also presented a qualitative comparison of the depth cameras employed by the system.

The next chapter will present two instances of BEAMING which have been developed during the first and third year of the project, respectively. The systems' architecture will be described in details, highlighting methodological contributions to the development, and presenting application scenarios to evaluate the platform with respect to typical telepresence properties such as spatiality and embodiment.

Chapter 4

BEAMING Platform Instances

Reality is merely an illusion, albeit a very persistent one.

Albert Einstein

This chapter introduces two instances of the BEAMING platform. The first instance, named "platform one" (BP1) henceforth, is the result of the first year of development. Section 4.1 details the specification of BP1 and introduces a case study, the acting scenario, through which the platform was evaluated. The second platform, termed "platform two" (BP2) henceforth, is the evolution of BP1 and is the result of the third year of development. Section 4.2 introduces BP2's architecture, describing a novel case study, the remote meeting, and highlighting improvements over platform one.

The BEAMING platforms presented here are a result of a collaborative development effort shared between a variety of institutions. Throughout the chapter then, methodological contributions made by the research documented in this thesis will be highlighted. With respect to BP1, the main contributions are related to video acquisition (both panoramic and 2.5D), rendering and streaming of the destination. In particular, a novel solution to stream video-plus-depth over networks was developed for this platform. Regarding platform two, the contributions mostly focus on acquiring 3D static geometry of the meeting room, calibrating this with a live, mesh-based reconstruction of the locals and calibrating cameras located at the destination with the static geometry. Please note that some of the images used in this chapter are adapted from the author's own work [SNO+12, PKW11].

4.1 BEAMING Platform One

This section describes the BEAMING platform one, and its use to support remote acting rehearsal. The rehearsals involved two actors, located in London and Barcelona, and a director in another location in London. This triadic audiovisual telecommunication was performed in a spatial and multimodal collaborative mixed reality environment based on the BEAMING's "destination-visitor" paradigm introduced in the previous chapter.

We will detail the heterogeneous system architecture, which spans the three distributed and technologically asymmetric sites, and features a range of capture, display, and transmission technologies. The actors' and director's experience of rehearsing a scene via the system are then discussed, exploring

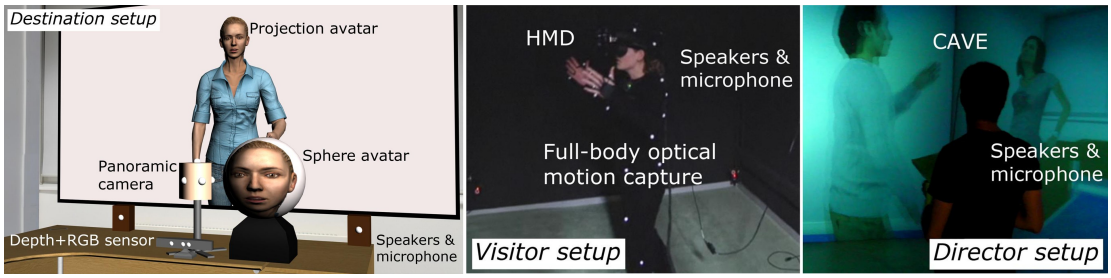


Figure 4.1: *Platform one: asymmetrical technical arrangements at the three sites.*

successes and failures of this heterogeneous form of telecollaboration.

Please note that throughout the platform, audio capturing, transmission and playback is performed using Skype [Mic02], stereo speakers and a microphones.

4.1.1 System Architecture

Figure 4.1 illustrates the distinct arrangements at each of the three sites in our particular studied setup: the physical destination, where one actor is located, is equipped with a range of capture and display technologies; the visitor site, at which the second actor is located, is composed of an immersive HMD-based VR system with full-body tracking; and the directors setup is an immersive CAVE-like system, although it could be a standard machine located anywhere. In this triadic interaction, the visiting actor and the director can be classified as visitors. However, the director plays the role of a spectator, as he/she is not visually represented to either actors, so that he/she can move around the shared space without causing distraction or occlusions to the actors.

Destination Site

The destination site is a physical meeting room, measuring approximately $5 \times 3 \times 3$ m. Following BEAMING's destination-visitor paradigm, the destination actor is physically located at the meeting room while the visiting actor is remotely perceived and virtually represented. The director, which in here acts as a spectator, is not visually present in the shared VE, but perceives the destination, including virtual representations of both actors. We have already argued that one of BEAMING's requirement is that technical interventions must be small and unintrusive. Hence, the main feature of the destination is that it should largely remain a standard meeting room to any collocated people, in the sense that they should not be encumbered by worn devices, such as wires or HMDs, while taking part in the action.

Acquisition. The destination must be equipped with technology able to acquire both the environment and the co-present actor to transmit to the visitor and director sites. To this aim, the platform supports two methods of visual capture of the local environment, and three methods of visual capture of the local actor. Figure 4.2 shows how all of these modes as they appear at the visitor site (these are discussed later in the section).

The first acquisition mode is performed using omnidirectional capture. Spherical video acquisition, which implicitly captures both actor and environment, is achieved with a Point Grey Research Ladybug 3 camera (see Section 3.2.2 for hardware specifications). The camera provides a simple means of visually capturing the destination and the people within. This surrounding acquisition is directly compatible with

Figure 4.2: *The three display modes available at the visitor site. From left to right: spherical video, embedded 2.5D Kinect video of the destination actor within a VE, and a pure VE featuring Kinect-tracked embodied avatars.*

the immersive display characteristics at the visitor and director sites, both of which feature wide FoV displays.

The second method of environment capture is a hand-made 3D textured model of the destination room. The model's wall textures are re-projected from a 50 MP panorama, while the furniture textures are extracted from single photographs.

Omnidirectional capture of the destination also allows for real-time dynamic visual capture of the actor. The other two modes of capturing the destination are performed using a Microsoft's Kinect sensor (see Section 3.2.3 for hardware specifications). The first Kinect-based solution for capturing the destination actor makes use of the PrimeSenses' OpenNI and NITE (Natural Interaction Technology for End-user) middleware libraries [App10]. NITE allows for skeletal recognition and tracking using the depth-sensing abilities of the Kinect. While the tracking data are not as high-fidelity, high-frequency, or low-latency as a professional motion capture system, NITE has the significant advantages of being marker-less, requiring minimal technical setup and calibration time in-line with BEAMING's principles. The calculated skeletal data are transmitted to the visitor and destination site at 30 Hz, and are used to animate a graphical avatar representing the destination actor. The final mode of capturing the destination actor, and the second Kinect-based mode, streams a 2.5D point-based textured video representation of the actor, independent of the environment, at 30 Hz. This representation may be considered a 2.5D video avatar, as we only employ one front-facing Kinect to record the actor, and thus do not provide coverage of the rear half of the body. To extract the actor's body, we leverage knowledge from the NITE-based skeletal tracking and place a bounding box in the depth range at which the tracked joints are currently positioned. To stream the 2.5D video avatar, we developed a compression algorithm able to adapt conventional video codecs to depth streaming. Details of this technique are given in Section 4.1.1.

Display. Besides acquisition devices, the destination is equipped with display technology to represent the visiting actor in a manner that fosters a physical presence. To this end, the system architecture offers two solutions: a large high-resolution avatar projection (i.e. a PROJECTION-avatar) and a 360° spherical display showing an avatar head only (i.e. an ICON-avatar). Even if both solutions are not mobile, restricting the visitor's movement around the destination, each display type aims to provide a distinct benefit to the remote interaction.

The projection avatar, for instance, enables life-size and full-body embodiment of the visitor. Due to the corresponding full-body motion capture setup at the visitor site, an avatar representing the visitor

is puppeteered in real time. Thus, nuances of the visitor's body language are represented. Due to the large 3×2 m screen size, gross movement on the horizontal and vertical axes, and to a lesser extent on the forward-backward axis, are also supported. This means that if the visitor walks from left to right, sits down or jumps, then these movements will be clearly conveyed by the PROJECTION-avatar. The same cannot be said for movements such as getting closer or farther away from the virtual camera, as this will result either in increased or reduced size rather than being displayed at the physical position.

The second mode of visitor display is the use of an ICON-avatar. The display aims to foster a greater sense of presence of the visitor at the destination, and to help the locals understand his/her 360° directional attention. The avatar head as displayed on the display rotates and animates in real time based on head tracking, eye tracking, and voice-detection data acquired at the visitor site. Even though this representation makes impossible to represent visitor's movement, its usage in combination with the PROJECTION-avatar mitigates this limitation. Implementation details of the ICON-avatar, together with a user study, are detailed in Oyekoya *et al.* [OSS12].

Visitor Site

The technology at the visitor site is responsible for both capture of the visitor and the immersive display of the destination and its collocated locals. In the BP1 setup, this comprises of a VR facility at which the technologies for acquisition and display are a full-body motion capture system and an immersive HMD, respectively.

Acquisition. Capture of the visitor is performed using a NaturalPoint Optitrack [Nat96] motion capture system consisting of twelve cameras. As the capture volume ($3 \times 3 \times 2.5$ m) is smaller than the meeting room, a one-to-one mapping between the visitors movements in the perceived virtual destination, and the position of their embodiment at the physical destination is possible. The skeletal data available from the Optitrack has higher-fidelity than the equivalent Kinect NITE tracking at the destination, albeit the usage of a motion capture suit (see Figure 4.1) with a greater calibration time ($\sim$ 20 min as opposed to $<$5 min for NITE) is required for this solution. The data is then streamed to the destination with the protocols detailed in Section 4.1.1.

Display. Display of the destination and local actor to the visitor is achieved using a NVis nVisor SX111 HMD [NVI08]. The HMD features a 111° horizontal $\times$ 64° vertical FoV and a resolution of 1280×1024 displayed at 60 Hz. The visual modes captured and transmitted from the destination, which have been detailed in the previous section, may be dynamically swapped between. To improve spatial reference and presence, as suggested in Mohler *et al.* [MCRTB10], the platform associates a virtual avatar to the visitor also in his/her visualisation: in this way, the visitor can look down and see his/her own virtual body. VRMedia's XVR [TCB+10] software framework is used to render the VE; the avatars are rendered using the Hardware Accelerated Library for Character Animation (HALCA [GS10]). To correctly represent the 360° visual stimuli of the Ladybug3, we mapped the equirectangular projection texture of the destination to a sphere, effectively employing a spherical projection, and we used OpenGL [Khr00] to render it. This solution allows the visitor to look around the sphere as if he/she is looking around the destination.

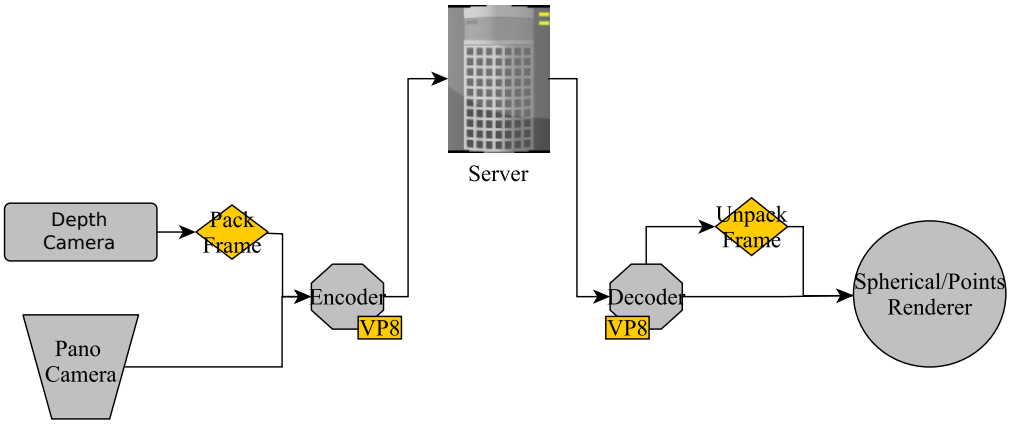


Figure 4.3: *Networking architecture for surrounding and colour-plus-depth streaming using VP8 codec. Yellow rhombi represent the depth encoding/decoding scheme developed in this research [PKW11].*

Director Site

Since the director is not represented visually, there are no acquisition devices at his site. Regarding the display, the director can use any device, ranging from non-immersive flat displays up to immersive systems. In this instance of the platform, the director was placed in a CAVE immersive system, where he could physically navigate the remote environment and, by shifting his view-point, naturally review the actors performance. Thus, the director views a VE of the destination, populated with the avatar embodiments of the two remote actors. However, given the asymmetry of motion capture systems between the destination and visitor sides, the destination actor may be perceived to have fewer degrees of freedom (e.g., NITE does not track head, wrist, and ankle orientation) than the visitor, and as a result may appear more rigid and, due to the lower capture rate, less dynamic.

Transmission

Communication between participants distributed over the three international sites relies on low-latency data transmission. As the nature of the various media streams originating at the sites can greatly vary, a monolithic transmission solution is inappropriate, and would likely result in network congestion. Hence, we divide the media into two types by bandwidth requirement: low and high bandwidth.

Low-bandwidth data comprise session management and skeletal motion capture data, and its transmission is handled by the BEAMING Scene Server (BSS) introduced in Section 3.1.3. To enable later playback, we enabled BSS's data logging capability.

High-bandwidth data are composed of video acquired from the Ladybug 3 and Kinect cameras at the destination site and transmitted to the visitor site for display in the HMD. Several solutions to encode, transmit and decode video streams, including the transmission of color-plus-depth data, were investigated. This resulted in the unified streaming architecture illustrated in Figure 4.3. The platform's end-to-end surrounding video transmission solution implements Google's VP8 encoding with RakNet streaming. Colour frames are grabbed from either a Ladybug3 camera or a Kinect; the latter also provides depth data. Subsequently, colour frame are passed to the VP8 encoder, while depth-maps are firstly

encoded with a novel encoding solution developed in this research [PKW11], and then sent to the VP8 encoder. As both depth and colour are encoded using the same encoder, we stream a colour-plus-depth compressed data as a single packet, and then reconstruct this on the receiver side with the same decoder. Packets are sent to the central server, and then relayed to the receiver which decodes and sends to the renderer both colour and depth data. Our implementation achieves frame rates of $\sim$ 13 Hz (from the original 15 Hz) for the Ladybug3 camera and $\sim$ 20 Hz (from the original 30 Hz) for the Kinect in the visitor's HMD, and end-to-end latency of transmitted frames is $<$ 200 ms.

Surrounding Video Streaming. The bandwidth generated by a Ladybug camera ($\sim$ 90 MB/sec) demands a high level of compression to achieve real-time transmission. For this reason, the Ladybug raw panoramic RGB images are firstly converted to YUV space, after which the YUV image is compressed using the Libvpx VP8 codec library [Web08, Goo11]. The compressed frames are then sent as a RakNet *bitstream*, a mechanism to compress and transmit generic raw data, to a server process (in our case located in Pisa, Italy), which simply relays the stream to other connected peers such as the visitor site running the HMD. Upon being received, the compressed VP8 frames are decompressed to YUV and then converted back to RGB space.

Depth-Enabled Video Streaming. While VP8 is the best solution for panoramic video streaming, the same cannot be said for streaming depth maps. Streaming the information available from depth cameras is non-trivial due to the type of data employed (16-bits in our case) and the required bandwidth ($\sim$ 18 MB/sec for a Kinect camera). In BEAMING we are interested in a general solution that adapts standard video codecs, such as Google's VP8 or H.264 [MWS06], to depth streaming, as this allows the platform to use a single streaming layer for a variety of data. To this end, we have developed a novel encoding scheme that efficiently converts the single-channel depth images to standard 8-bit three channel images, which can then be streamed using standard codecs. Our depth-map compression scheme is designed to be resilient to quantisation, and comparatively robust against down-sampling (convolution) and altered intensities due to lossy compression. Our encoding scheme ensures that the sent depth values are received and decoded with a high degree of accuracy. Figure 4.4 shows an overview of the method. A detailed description of the method, together with a discussion and evaluation using different video codecs, is reported in Appendix C.

Our solution works as follows. We express our scheme as a mapping from integer depth values $d \in \{0,\ldots,w-1\}$ ($w = 2^{16}$ for a 16-bit depth map) to three $[0,1]$-normalised (colour) channels $L(d)$, $H_a(d)$ and $H_b(d)$. $L(d)$ is a linear mapping of d into $[0,1]$ and, since subject to quantisation, is interpreted as a low-depth-resolution representation of d,

$$L(d) = (d + 1/2) \,/\, w \,,$$

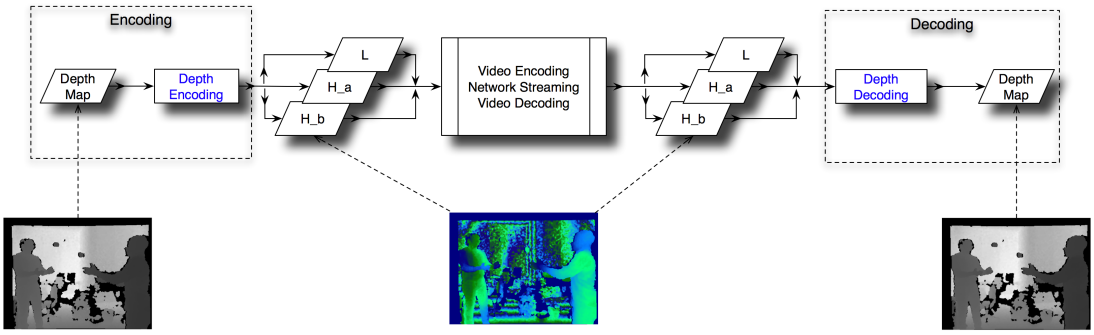


Figure 4.4: *Graphical overview of the proposed method. The original 16-bit depth map is encoded in an 8-bit, three-channel image and is then processed by a video encoder and transferred over the network. When received, the three-channel image is decoded through the video decoder and is then processed by our method to reconstruct the original 16-bit depth map.*

while H_a and H_b are chosen as fast-changing, piece-wise linear functions (triangle waves) whose slopes are high enough to be expressed in the low-precision output representation:

$$H_a(d) = \begin{cases} \left(\frac{L(d)}{\frac{p}{2}} \bmod 2\right) & \text{if } \left(\frac{L(d)}{\frac{p}{2}} \bmod 2\right) \leq 1 \\ 2 - \left(\frac{L(d)}{\frac{p}{2}} \bmod 2\right) & \text{otherwise .} \end{cases}$$

$$H_b(d) = \begin{cases} \left(\frac{L(d)-\frac{p}{4}}{\frac{p}{2}} \bmod 2\right) & \text{if } \left(\frac{L(d)-\frac{p}{4}}{\frac{p}{2}} \bmod 2\right) \leq 1 \\ 2 - \left(\frac{L(d)-\frac{p}{4}}{\frac{p}{2}} \bmod 2\right) & \text{otherwise .} \end{cases}$$

n_p is the integer period of H_a and H_b in the input depth domain and needs to be at most twice the number of output quantisation levels ($n_p \leq 512$ for 8-bit output); $p = \frac{n_p}{w}$ is this period normalised to a $0 \ldots 1$ depth range. Thus designed to be resilient to quantisation, H_a and H_b will be used to decode fine-grain depth variations, while L will anchor these variations in the global depth frame.

In practice, $L(d)$, $H_a(d)$ and $H_b(d)$ can be tabulated for any d in the input depth range, reducing depth encoding to a simple look-up with negligible computational overhead. H_a and H_b are triangle waves with equal period and different phase. The phases are chosen, so that for any depth value $\bar{d}$ encoded by L, either H_a or H_b is linear within $\bar{d} \pm p/4$.

Once the original depth data is split in a triplet $(\bar{L}, \bar{H_a}, \bar{H_b})$, these values can be encoded, streamed and decoded with any video codec. In our implementation we use Google's VP8 codec, and details of this are given in Appendix C.

Accordingly, given an encoded triplet $(\bar{L}, \bar{H_a}, \bar{H_b})$, the original depth value $\bar{d}$ can be decoded by

determining a depth offset L_0 from L and adding a fine-scale depth correction δ:

$$\bar{d}(\bar{L},\bar{H}_a,\bar{H}_b) = w \cdot \left[L_0(\bar{L}) + \delta(\bar{L},\bar{H}_a,\bar{H}_b)\right] ,$$

$$\delta(\bar{L},\bar{H}_a,\bar{H}_b) = \begin{cases} \frac{p}{2}\bar{H}_a & \text{if } m(\bar{L}) = 0 \\ \frac{p}{2}\bar{H}_b & \text{if } m(\bar{L}) = 1 \\ \frac{p}{2}(1-\bar{H}_a) & \text{if } m(\bar{L}) = 2 \\ \frac{p}{2}(1-\bar{H}_b) & \text{if } m(\bar{L}) = 3 \end{cases}$$

with

$$L_0(\bar{L}) = \bar{L} - \left(\bar{L} - \frac{p}{8} \bmod p\right) + \frac{p}{4}m(\bar{L}) - \frac{p}{8}$$

$$m(\bar{L}) = \left\lfloor 4\frac{L(\bar{d})}{p} - 0.5 \right\rfloor \bmod 4 .$$

H_a and H_b are chosen to be triangle waves to be robust against spatial filtering; alternative choices, such as a saw-tooth wave, would have suffered from strong distortions at their discontinuities. While other mappings may still be possible, we argue that C^0 continuity is a desirable property, in particular where the codec downsamples individual colour channels.

4.1.2 Contribution

The candidate's main contributions to the BP1 is in the acquisition and transmission of the destination to the visitor. Specifically, the candidate has been the main developer of the surrounding video and 2.5D video acquisition, rendering and transmission solutions. To this end, he has developed solutions to interface with the cameras, render their video streams and transmit the data over networks. In particular, the work conducted for the depth-enabled streaming of the locals'avatar resulted in a peer-reviewed scientific paper [PKW11], of which he is the leading author.

4.1.3 Case Study: Acting Rehearsal

To test our platform, we hired three experienced theatre actors/directors to take part in a rehearsal, which took place over a period of 4 hours in a single afternoon. The participants, some of which were members of the Royal Academy of Dramatic Arts (RADA)[1], were paid £60 each to take part in the rehearsal. Prior to the rehearsal, the actors had, separately and apart, learned the "spider in the bathroom" scene from Woody Allen's 1977 movie *Annie Hall*[2]. The scene begins when Alvy, played by Woody Allen, receives an emergency phone call (actually a false, manufactured crisis) to come to Annie's (played by Diane Keaton) apartment in the middle of the night. He arrives and an hysterical Annie wants to be rescued from a big spider in her bathroom. Initially disgusted ("Dont you have a can of Raid in the house? I told you a thousand times. You should always keep a lot of insect spray. You never know who's gonna crawl over."), Alvy skirts around the issue for 2-3 minutes; firstly by discussing a rock concert program on Annie's bureau, and then a National Review magazine that he finds on her coffee table.

[1] https://www.rada.ac.uk

[2] An extract from the scene can be found here: http://youtu.be/OX5BngxRWLg. Accessed January 29, 2014.

Figure 4.5: *The acting rehearsal in progress at each of the three sites.* Left*: The destination site at UCL.* Center*: The visual stimuli (running in VE/Avatar display mode) of the destination site and actor displayed in the HMD at UB.* Right*: The director located in UCL's CAVE.*

An arachnophobe himself, Alvy eventually goes on to thrash around in the bathroom, using Annie's tennis racket as a swatter, in an attempt to kill the spider: "Dont worry!" he calls from the bathroom, amidst the clatter of articles being knocked off from a shelf. This scene was chosen as it consists of varied spatial and interpersonal interplay between the two characters. Thus, the actors engage in intense talk on varied subjects, spatial action (particularly Alvy's character), and directing attention toward and manipulation of objects in the environment. The scene's duration is 3 minutes in the original movie. This short length allows for multiple run-throughs over the 4-hours rehearsal period, and encourages the director and actors to experiment with new ideas and methods toward the final performance.

The male character, *Alvy*, was portrayed by a male actor at the destination site at UCL, while the female character, *Annie*, was played by an actress at the visitor site at UB. The male director was located in UCL's CAVE facility, separate from the destination room. [3]. Figure 4.5 shows the three distinct views of the rehearsal space.

Following the rehearsal, we had a discussion with the actors and director, together with a group of three experienced theatre artists and academics who were spectators at the two UCL sites. During the discussion we recorded on paper the comments made by both the participants and spectators. The discussion was conducted as an informal interview, with no written questions to answer, and the notes were collected by Dr. William Steptoe and by the candidate. Drawing from these notes and anecdotal facts noted during the experiment, following we discuss the successes and failures of the rehearsals in terms of the central elements of spatiality and embodiment.

Spatiality. The common spatial frame of reference experienced by all parties was highly conducive to the nature of theatrical acting and directing. The artists were able to perform blocking (i.e. the precise movement and staging of actors on a space), referring to their movement and positioning with relative ease. This was demonstrated through the director issuing both absolute and relative instructions interchangeably. For instance, asking Alvy either to pick up the magazine "on the table" or "to Annie's right" were both unambiguous to all parties due to the aligned visual environment. The director was able to issue such blocking instructions on both macro and micro scales, ranging from general positioning and Alvy's point of entrance into the scene, down to the technical aspects of movement on a per-line basis.

[3]The male character was portrayed by Jannik Kuczynski, while the female character was played by Jasmina Zuazaga. The director was Morgan Rhys.

Figure 4.6: *Scenes from the virtual rehearsal.*

It's important to note that the artists considered the asymmetry in allowed range of physical movement between the two actors, based on their status as a local or a visitor, as a limitation. The destination actor was free to move around the entire rehearsal space, and would be observed by the visiting actress and director as doing so. However, the visiting actress' allowed movement was limited, particularly forward and backward, due to her situated representation at the destination (projection ans sphere avatars), which greatly differ in terms of how well they accurately represent the position of the visiting actress. In particular, the projection avatar display, which covers the whole rear wall of the destination site, is able to represent horizontal and vertical movement well. Depth cues, however, are less easily perceived, and a forward movement performed by the visiting actress results in the projection avatar getting larger, due to being closer to the virtual camera, rather than having physical presence at a location further into the destination room. The sphere avatar display only shows an avatar head representation, and so cannot express bodily movement at all. The implications of this differing movement allowance between the two actors resulted in some frustrations for the director, who reacted by issuing more gross blocking instructions to Alvy, while focusing more on instructing Annie's expressive gestures. This situation, however, matched the scene's dynamics, in which Alvy is the more physically active of the two characters. Figure 4.6 illustrates some key moments during the scene, captured with our virtual replay tool.

The solution to this issue of the visitor being spatially restricted at the destination is the use of mobile displays. However, the use of personal telepresence robots is likely to solve one set of issues, to the detriment of others. On one hand, they would provide a physical entity which could freely move in space. However, the predominant design of such devices is a face-only LCD display, recorded from webcam video. Thus, the fuller body language and gestural ability provided by the avatar representation of the visitor would be missing, which is a critical cue used while communicating [Duc86]. Some general observations on the benefit of the common frame of reference were also made by the spectators and participants. For instance, our senior guest academic, Edward Kemp - the Artistic Director of RADA, discussed the way that many actors are able to learn their lines more quickly by physically being in the rehearsal room or theatrical set as opposed to being in a neutral location such as their own home. In particular, some older actors can only learn lines once they have established the blocking of a scene. Hence, the interactive and visual nature of the system was considered highly beneficial to the process of learning lines and planning movements, even in a solo rehearsal setting.

Embodiment. The interactions were significantly influenced by modes of embodiment and display at each site. Firstly, it should be noted that throughout the rehearsal period, no critical failures in communication occurred. While we have not formally measured the end-to-end latency of all modes of capture,

transmission, and display, this suggests that it is acceptable to support both the verbal and non-verbal triadic interaction. The initial period of acclimatization to the interaction paradigm resulted in some confusion between the three participants due to the evident asymmetry between them. Each party was unclear about the nature of the visual stimuli the others were perceiving. Once some initial descriptions were provided by each party (the destination actor only needed to provide minimal information as he was physically present in the place where the others were virtually present), the group became confident about the unified space they were all perceptually sharing, together with their displayed embodiments.

The initial period of the rehearsal was used to determine each participant's local display preferences. At the destination site, the projection avatar was preferred over the sphere avatar or a dual-representation of the visiting actress. The destination actor considered the projection avatar to provide more useful information through the display of full-body language as opposed to the attentional cues that the head-only sphere display provided. Simultaneous use of the projection and sphere avatars was disliked as it resulted in confusion due to division of attention between two locations. The projection avatar bestowed Annie with a higher degree of physical presence for Alvy to play against and observe (which enhanced the physicality of the performance from Alvy's perspective). During the post-rehearsal discussion, Alvy recalled his excitement when Annie took a step toward him, and an impression of their close proximity was provided by the depth-cue of Annie's avatar increasing in size on the projection display: *"When you do go close to the screen; when there are situations where you're flirting, when she's supposed to touch my chest and so on, that is really interesting because she's in Barcelona and I'm here, but there's still some part of you that tries to reach out and touch her hand on the screen. And when she reacts; for instance when I start smacking the floor looking for that spider, she automatically did that [gestures to cover his head]sort of thing. There was interplay between us; a natural reaction to what I was doing. That was exciting and when the project shined the most, in my eyes."*

The visiting actress wearing the HMD decided to observe the destination using the panoramic video mode as captured by the Ladybug 3. This mode preserved the actual appearance of both the destination and Alvy, with the trade-off being a decreased perception of depth due to the monoscopic video. This mode was preferred over both the VE/Avatar and VE/2.5D video display modes, due to the improved dynamism of the video compared to the "stiff" avatar embodiment that did not feature emotional facial expression, and the clearer image of Alvy due to the higher resolution camera.

Central to the system's asymmetry are the physical abilities of the two actors depending on at which site they are located. There are several moments during the scene when the actors are required to interact with each other and their environment. This includes knocking on a door; looking at, picking up, and passing objects; and hitting an imaginary spider. When performing such actions during the rehearsal, Alvy has a tangible sense of doing so due to the physicality of his local environment. So, when he knocks on the door or picks up a magazine, he is doing just that, and these actions (and sounds) are observed by both Annie and the director, albeit in varying visual forms. However, this ability does not extend to Annie, as, regardless of display mode, the visitor is only able to mime interaction with perceived objects that are, in reality, located at the destination. Fortunately, most of these moments in the scene belong to

Alvy rather than Annie, so this issue did not result in critical failures.

The director in the CAVE viewed the rehearsals as a VE populated with the two actors' virtual avatars. Due to Annie's actual appearance not being captured by video cameras, an avatar is her only available mode of representation at the other sites. While both video and avatar representations of Alvy are available in the CAVE, the director preferred visual consistency, and preferred the avatar representations in the VE. He also considered his ability to freely move and observe the actors within the rehearsal space as a powerful feature of the system. He was able to observe the scene from any viewpoint, which allowed him to move up close to the actors to instruct the expressive dynamics of their relationship, or stand back and observe their positions in the scene as a whole. The fact that the actors were represented as life-sized avatars aided direction by enhancing the interpersonal realism of the rehearsal. Both actors noted our decision to not visually represent the director. Although the benefit of the director's unobstructive movements was universally acknowledged, the inability of the director to use non-verbal gesture, particularly pointing, was considered a hindrance to the rehearsal process. Allowing the director to make his representation visible or invisible to the actors is a potentially interesting avenue of investigation that may have implications for general telecommunication in such systems.

The overall impression of the abilities afforded by the actors' embodiments over the three sites was that movement and general intent was communicated well, but details of expressive behaviour were lacking. Facial expresion, gaze, and finger movement were highlighted as the key missing features. (The BP1 is able to track, transmit, and represent gaze and finger movement with high fidelity, and some facial expression is supported; however, these cues require participants to wear encumbering devices, and so were decided against for this rehearsal application.) As a result, moments in the scene that have intended emotional prescience, such as those featuring flirting, fear, and touch, appeared flat. In an attempt to counter these limitations of expressive ability, the actors noted that their natural (and at times subconscious) reaction was to over-act in order to elicit a response from their partner. Correspondingly, the director found himself requesting the actors to perform exaggerated gestures and movements that he would not have done if the finer facets of facial expression were available.

Discussion. Depending on the characteristics of the play or production, the artists speculated that rehearsing using the BP1 could reduce the subsequent required collocated rehearsal time by up to 25%. The primary benefit to the rehearsal would be blocking the scene, planning actors' major bodily gestures, and, in the case of television and film work, planning camera shots and movement. In television and film work, the artists noted that rehearsal is often minimal or nonexistent due to time and travel constraints. The system provides a potentially cheaper and less time-consuming mode of being able to rehearse with remote colleagues. This benefit would likely extend to technical operators and set designers, who would be able to familiarize themselves with the space in order to identify locations for technical equipment, and optimize lighting and prop-placement. The heterogeneity and multimodal nature of the BP1 was also suggested as a novel paradigm for live performance in its own right, including the potential for art and science exhibitions, and even reality television.

Blocking and spatial dimensions are paramount to a theatrical scene, and determining these aspects is frequently divisive between international performers. Such disputes may be reduced or settled early by allowing all parties to virtually observe and experience the rehearsal or set layout prior to a collocated performance. Both actors and the director advocated the system as a means of overcoming the initial apprehension and nervousness of working with one another, and suggested that they would be more immediately comfortable when the time came for a subsequent collocated meeting. Solo performance and reviewing prior run-throughs was discussed as a potentially useful mode of system operation. To this end, the virtual replay abilities of the system allow for random-access and time-dilation of previous sessions.

A key strength of the platform is its ability for remote participants to move within and observe a perceptually unified space. This aspect of the system was often exploited by both the actors and the direct, as they made full use of the virtual space they were given. It's important to note that the relative inexpressivity of the actors' embodiments implies that scenes relying on performing and reacting to consequential facial expression and subtle gesture would not benefit significantly from rehearing via the BP1 in its current form. Hence, to partially overcome this limitation, in the BP2 we introduced more fine-grained gesture and facial-expression tracking systems.

4.2 BEAMING Platform Two

The BP1 demonstrates some of the key aspects of BEAMING. One of the biggest outcome of the platform is that asymmetry of technologies can indeed support symmetry of virtual experience. However, the BP1 was an early instance of the BEAMING idea, and as such lacked some of its key functionalities. Therefore, with the BP2, we further developed the initial platform by extending the capture and display modalities to match the initial requirements.

In this section we will present the BP2, and its use to support remote meetings. Following the BEAMING's destination-visitor paradigm, the meeting involved a physical meeting room with co-located locals and a remote visitor, both located in Pisa. We will detail the heterogeneous system architecture, which spans the two distributed and technologically asymmetric sites, and features a range of capture, display, and transmission technologies. The experience of the remote meeting via the system is also discussed.

4.2.1 System Architecture

Figure 4.7 illustrates the distinct arrangements at each of the two sites in our particular studied setup: the physical destination, where multiple locals are located, is equipped with a range of capture and display technologies; the visitor site is an immersive CAVE-like system. In this interaction, while the visitor is restricted to be a single user, the number of the locals supported has, theoretically, no limit. In practice, the number of locals supported is limited by the FoV of the visual acquisition devices.

Destination Site

The destination site is a physical meeting room, measuring approximately $5 \times 8 \times 3$ m. Following BEAMING's destination-visitor paradigm, the destination locals are physically located at the meeting

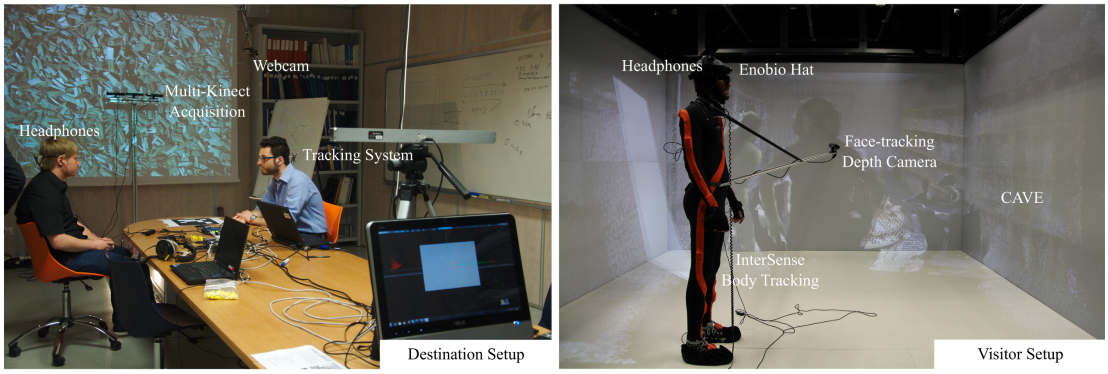


Figure 4.7: *Platform two: asymmetrical technical arrangements at the two sites.*

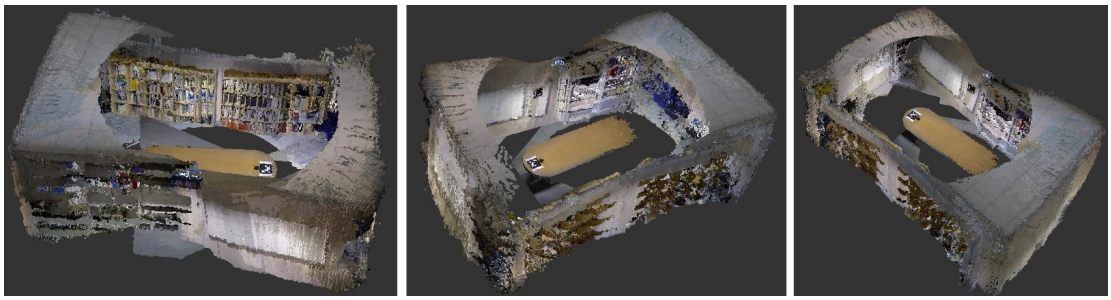


Figure 4.8: *3D model of the destination acquired with the RGBD-mapper.*

room while the visitor is remotely perceived and virtually represented. In line with BEAMING's requirements, technical interventions at the destination are small and unintrusive. To this end, the destination largely remains a standard meeting room to all the collocated people, as the hardware required to run the platform is unintrusive and minimal. Additionally, no locals are required to wear HMDs or tracking suits, but only headphones if binaural audio is enabled.

Acquisition. The destination is equipped with technology able to acquire both the environment and the co-present locals to transmit to the visitor site. To this end, the platform supports a method of acquiring the local environment, and a method of visual capture of the locals.

The local environment is acquired using a RGBD-mapper (see Section 2.6.2 for works related to the topic). A RGBD-mapper is a solution to gradually build up a point-based 3D model from a RGBD video stream. The system facilitates the preparation of a physical destination for BEAMING as it automates the acquisition of a static 3D model of the environment. The system improves the manual creation of models from panoramic imagery employed in the BP1. In contrast to the aforementioned approach, the RGBD-mapper allows rapid acquisition, also of cluttered environments with many depth discontinuities, that would otherwise be very hard to manually model from panoramic image material. For the current implementation of the platform we employed a modified version of *RGBDSLAM* [EHE+12]. Figure 4.8 shows a point cloud of the meeting room acquired with the RGBD-mapper.

The mapper works as follows: a depth-camera (a ASUS Xtion PRO Live unit for the results showed here) is swivel-mounted on a tripod and is placed centrally in the room that needs to be modelled. To start the reconstruction, the user only needs to rotate the camera around its vertical axes while a real-time

feedback of the model is given on screen. When the room is entirely mapped, the model can be exported in one of the available formats (i.e. ply, pcd or ptx). We modified the system to support some additional features. For instance, the user can choose to set a filter to tune the point density of the final model or insert external scans in the environment. To insert external scans in the environment, the system relies on the AR tracking library ArUco [GJMSMCMJ14]. Each scan needs to be accompanied by the initial RGB frame employed for the modelling with a specific AR marker in view. This allows ArUco to calibrate the various scans, by finding a calibration matrix for each model that places the 3D marker location in the origin of the common coordinate system. Once the models are acquired and calibrated, the final output is uploaded to a central server from where it can be downloaded by the transporter's software.

Visual capture of the locals is managed with a free-viewpoint, multi-depth-camera based reconstruction solution. In the BP1 locals were captured using algorithms to fuse geometry data from multiple depth or surrounding video reconstructions and suitable compression and transmission of depth data. In the BP2 these solutions culminated in end-to-end capabilities for free-viewpoint rendering that allow capturing, transmission and display of a high-quality 3D model of a subject. The key design aspect of this solution is that rather than reconstructing a single 3D model before transmission to the visitor, we send partial lower-level reconstructions to the transporter, performing the final visual reconstruction at the visitors site. This strategy has two main advantages: a) transmission of lower-level reconstructions allows for more time and space efficient encoding and b) display-side fusion of the partial reconstructions allows the renderer to cull data sources that do not fall into the visitors current scope.

The free-viewpoint solution works as follows: multiple ASUS Xtion PRO Live cameras are placed in the meeting room so that they cover a significant area of interest. In our setup three cameras were placed on top of a T-shaped stand, one next to each other. Internal calibration of this camera network is performed prior to the meeting. Additionally, the cameras are also calibrated against the 3D model acquired with the RGBD-mapper using a marker-based strategy. Each depth map retrieved from the various units is coarsely meshed using a meshification algorithm which produces a 3D triangle mesh which approximates the original (or background-segmented) depth map (see Figure 4.9). The algorithm initially selects a set of seed points on the depth image in correspondence of depth discontinuities and non-planar surfaces. Delaunay triangulation is then performed on the seed points, thus producing a 2D triangle mesh in the XY plane which covers the convex hull of the captured surfaces. Subsequently, the 2D mesh is refined by splitting each triangle which spans over a depth discontinuity or over invalid depth regions. Finally, the mesh is extruded along the third dimension Z by assigning the corresponding depth value to each vertex and calculating the actual world coordinates for each mesh, thus producing the final 3D triangle mesh. The mesh is then compressed with the compression strategy introduced in Section 4.2.1, and streamed to the visitor site, where the various meshified depth-maps, as well as the 3D model of the room, are merged together using calibration information. Fuller details on the meshification algorithm are given in Bannò *et al.* [BGTB12].

In addition to the multi-depth-camera solution, compact webcams are placed in the room and streamed to the visitor site. The cameras are calibrated with the 3D models by using the same AR-based

Figure 4.9: *Multi-depth-camera based reconstruction results from three ASUS Xtion PRO Live cameras as rendered in the CAVE. Live dynamics of the locals are reconstructed and embedded within the destination's static geometry. Doubling of the image is due to the CAVE stereo renderer.*

calibration solution used to integrate multiple scans. Finally, as the platform supports binaural audio for the visitor, position of the locals is tracked using a Optitrack V120 Trio motion tracking system, while audio is captured using head-mounted microphones.

Display. Besides acquisition devices, the destination is equipped with display technology to represent the visitor to foster his/her physical presence. To this end, displaying of the visitor is performed using an AR-based solution (i.e. AR-avatar - see Figure 4.10) that runs on tablet devices. Each local is equipped with a tablet. The solution embeds a 3D avatar of the visitor with the view from the rear-mounted camera of the tablet. The avatar is puppeteered using tracking data sent from the visitor site, and it is embedded on the camera view using a feature-based camera tracking solution developed using the Qualcomm Vuforia toolkit [QT11] and Unity 3 rendering engine [Uni05]. The avatar animates in real time based on body tracking, eye tracking and emotion tracking data acquired at the visitor site.

The AR-avatar solution enables the locals to freely move in space and still be able to see the visitor from their correct view-points. Locals are not required to wear any particular device or stand in a specific place to correctly see the visitor. Additionally, the viewer can be run on fixed screen to allow a "virtual window" into the visitor site for spectators at the destination.

Figure 4.10: *AR-Avatar:* Left: *viewed using an iPad.* Right: *an avatar is embedded in the destination.*

Visitor Site

The technology at the visitor site is responsible for both capture of the visitor and the immersive display of the destination and its collocated locals. In the BP2's setup, this comprises of a VR facility at which the technologies for acquisition are a full-body and face motion capture system and an emotion capture system while the display facility is a CAVE system. Audio is captured using a head-mounted microphone, and played-back using stereo headphones and binaural audio.

Acquisition. Capture of the visitor is performed using a InterSense IS900 [Int96] acoustic motion capture system. While this solution provides accurate tracking, similarly to the BP1, it constrains the visitor to wear a motion capture suit. In addition to body tracking, visitor's facial expressions are captured using Faceshift [Fac12] capture system and an ASUS Xtion PRO Live camera mounted at $\sim$ 1m from the visitor's face. Emotional state capture is also performed through Enobio sensory capture system [Sta11] (see Figure 4.7). The data is then streamed to the destination with the protocols detailed in Section 4.1.1.

Display. Display of the destination and locals to the visitor is achieved using the $4 \times 4 \times 2.5$ m CAVE in Pisa. The visual modes captured and transmitted from the destination, which have been detailed in the previous section, allow the visitor to interact with the locals. VRMedia's XVR [TCB+10] software framework is used to render the VE. We employed OpenGL ES [Khr03] point-based graphics to render

the static point-cloud. To allow for high frame-rate, rendering is performed using GPU shaders and Vertex Array Objects (VBOs). As the model is statically rendered, the entire geometry is loaded on the GPU using VBOs. Then, frustum culling based on the visitor viewpoint is performed on GPU using dedicated shaders, to easy the number of points to render and speed the rendering process.

A similar process is used for the dynamic locals reconstruction streamed from the destination site. Here, we use OpenGL ES facilities to render triangles and cull objects which are outside the view frustum. Shaders are used to compute texture coordinates necessary to add colours to the models, while VBOs, with dynamic option enabled, are used to load the reconstruction data on the GPU. The same technique is used to render the webcam videos, for which we use billboards (i.e. OpenGL *Quads*) textured with the cameras' stream.

Transmission

Similarly to the BP1, communication between participants distributed over the two sites relies on low-latency data transmission. As the nature of the various media streams originating at the sites can greatly vary, we once again opted for a solution that divides the media into two types by bandwidth requirement: low and high bandwidth.

Low-bandwidth data comprise session management, skeletal motion capture data, emotion capture data and face performance capture data, and its transmission is handled by the BSS. High-bandwidth data comprises of videos from the webcam and geometry from the free-viewpoint meshification reconstruction. To this end, we developed a dedicated streaming solution based on Raknet's *bitstream* and VP8 encoding. Video streaming is performed in the conventional way, by compressing and decompressing each video frame using VP8 codec, and using UDP streaming. Compression and streaming of the 3D data is performed using a novel algorithm we developed. Such algorithm is a variant of a single-rate geometry compression algorithm based on the Valence Based Encoding [TG98] which is highly optimised to stream the triangle based reconstruction of the destination. Details of the solution are given in Bannò *et al.* [BGTB12]. The static 3D models of the destination are uploaded to a server and fetched by the transporter application prior to the meeting. Our implementation achieves frame rates of ~ 25 Hz (from the original 30 Hz) for each Kinect unit, and end-to-end latency of transmitted frames is < 200 ms.

4.2.2 Contribution

The candidate's main contribution to the BP2 is in the acquisition, transmission and rendering of the destination to the visitor. To this end, he has extended the *RGBDSLAM* mapper to allow for multiple models merging and filtering, and he has developed solutions to efficiently render the large point cloud generating from the mapper at the visitor site, which include dynamic frustum culling on GPU. In addition, the candidate has developed solutions to stream and calibrate a network of multiple webcams and he has contributed to solutions to calibrate the meshification reconstruction, 3D static models and video streams together.

Figure 4.11: *The virtual meeting in progress at each of the three sites.* Top*: The destination site with two locals.* Bottom*: The visitor located in Pisas CAVE.*

4.2.3 Platform Technical Test: Remote Meeting

To test our platform, we performed a remote meeting where several locals met with a single visitor using the BP2. Unlike the evaluation performed for the BP1 (cf. Section 4.1.3), we did not formally record any impressions or conducted interviews following the test. During the meeting, in fact, we were mainly interested in testing the various modules of the platform and demonstrating the functionalities of the system to a team of experts which were called to asses the status of the BEAMING project. Hence, in the rest of this section we will only report the successes and failures of the system from a technical point of view.

To demonstrate the various features of our system, we decided to organise a virtual meeting, during which all the features of the BP2 could be illustrated and tested. Hence, the participants of the meeting were instructed to discuss the BP2's architecture using a whiteboard and other objects, such as printed documents, located at the destination. As we were interested in demonstrating the feasibility of casual interactions between locals and visitors using our system, we did not specify a formal agenda for the meeting, but rather we asked the visitor to lead discussion. Additionally, we organised the meeting as an open meeting, in the sense that locals could casually join in or leave the meeting room during the entire length of the test (which run for one hour). This was an important test for our platform, as it helped demonstrating the fact that the BP2 does not require user instrumentation and can support a variety of modes, all with different level of fidelity. Therefore, while the initial locals could benefit from spatial audio, the users that joined the conversation as the meeting progressed did not, but this did not prevent them to take part in the discussion.

Most of the locals in the meeting were members of the PERCRO lab, the location in which the meeting took place. Additionally, four of the locals that joined the meeting belonged to the team of experts appointed to asses the project state. None of the locals had previous knowledge of the platform details. On the contrary, the visitor was one of the members of the platform development team (Sameer Kishore), and therefore he had in depth-knowledge of the system and was chosen as the leader of the meeting.

Figures 4.9, 4.10 and 4.11 show moments of the meeting, with different users acting as the locals. During the entire length of the test, the system run smoothly and this facilitated the communication between participants. The visitor extensively discussed with the locals technical details of the BP2, describing aspects related to the system's architecture and hardware as well as answering specific questions on the platform. In one occasion one of the locals drawn a sketch on the whiteboard to clarify one aspect of the streaming layer that was being illustrated by the visitor. As he was drawing, another local re-configured one of the webcams to capture the whiteboard and show it to the visitor, who could then comment on the drawing.

As already discussed, during the test we were mainly interested in analysing the technical performance of our system. With this goal in mind, we can conclude that our platform response to the test was satisfactory. No failures occurred for any of the modules of the platform, and this facilitated the discussion between participants, which run seamlessly and without interruption for the entire length of the test. Possibly the most interesting outcome of the meeting is that our system successfully handled a variety of different locals, with seamless switch between participants and scaling across an increasing number of users. Following the meeting most of the locals praised the usage of the AR-avatar display, which helped them in localising the visitor in the real space, and facilitated the interaction with him. Similarly, the visitor could greatly benefit from the dynamic reconstruction of users, as he could directly address new locals joining the conversation without any interruption in the discussion. Another feature that was positively perceived by the users was the possibility to dynamically reconfigure the webcams at the destination. Hence, from a technical point of view, we were satisfied by the test result, as it demonstrates how the BP2 improves on the BP1 in terms of the fidelity of visual, haptics and audio reconstruction conveyed.

During the test we also noted a variety of users' actions that suggest that users could successfully mimic behaviours which are typical of face-to-face interaction, such as spatial "spatial deixis". Spatial deixis is the reference by means of words (such as "this/that", "here", "next to") and/or gestures (such as pointing or gaze direction) that are dependent on context for their interpretation [Fil82]. In order for spatial deixis to communicate successfully, interlocutors need to have visual access to a common context, and share (or be able to interpret) one another's visual perspective [SCKMB03].

Interestingly, some of the users' behaviour noted during the virtual meeting confirmed what already discussed for the BP1. While presenting the system architecture, the visitor often made spatial references to specific areas of the destination and objects within. This was done by using a combination of pointing and verbal-cues, which suggests that the visitor could translate between his real space and the shared

destination space. This behaviour was mainly noted while discussing hardware components; in that occasions, the visitor often pointed to areas in the CAVE in which the objects appeared, and used words such as “next to this chair” or “here on the table”. Similarly, the locals often made spatial references to objects at the destination while discussing with the visitor. Typically, locals used words such as “to the left” or “in front of” to describe objects position to the visitors, which however had no problems in understanding. This suggests that both sides of the meeting had established a shared spatial reference, and, similarly to face-to-face communication, used it extensively to facilitate the discussion.

4.3 Chapter Summary

This chapter introduced the reader to two instances of BEAMING. It complements and expands the concepts introduced in Chapter 3, demonstrating successful implementation of the original project’s vision. Section 4.1 described the BEAMING platform one, the initial platform that was developed and tested after the first year of development. Technical details, including system architecture and dedicated hardware, have been introduced and discussed. In addition, a case study is introduced to test and evaluate the platform in terms of fundamental VE properties such as spatiality and embodiment. Section 4.2 introduced the BEAMING platform two, the evolution of the BP1. As for the previous system, technical details, including hardware and system architecture are introduced and discussed. A test scenario is then presented, with a short discussion on the test outcome.

The last two chapters have framed part of the research I have conducted during my studies, and prepare the ground for the following experimental chapter. Specifically, the next chapter will introduce a system, called *PanoInserts*, which can be considered as a particular instance of the BEAMING platform. The system, which we developed and enables practical spatial videoconferencing through portable devices, has been the main tool to perform my first investigation on whether videos in panoramic contexts can help users remotely perform collaborative, spatially demanding tasks.

Chapter 5

Experiment: Videos in Context for Telecommunication

Without knowing how to do it, I began to record some facts around me, and the more I looked the more the panorama unfolded.

Frederic Remington

This chapter presents an experiment designed to evaluate the impact of videos in panoramic context on remote, collaborative tasks that require a high level of spatial reasoning. As such, this chapter addresses one of the main questions presented at the beginning of this thesis. The previous chapter suggested that a technically asymmetric ICVE system such as BEAMING can benefit remote collaboration by presenting virtual shared spaces that users can intuitively understand and act upon. Starting from this finding, the aim of the study presented here is to understand if consumer devices, such as smartphones and tablet computers, can offer a similar experience. Specifically, we are interested in understanding if videos available from portable devices can be combined and represented in a way that offers enough information about the dynamics of remote places, supporting teleconferencing while achieving spatiality.

To support the study, we developed a teleconferencing system that uses smartphone cameras to create a surround representation of meeting places. We call this system *PanoInserts*. PanoInserts can be considered as a lightweight instance of the BEAMING platform, as it implements a network of commonly-available devices to achieve surrounding video conferencing for small-group interaction. Broadly speaking, PanoInserts works as follows: we take a static panoramic image of a location into which we insert live videos from smartphones. We use a combination of marker- and image-based tracking to position the video inserts within the panorama, and transmit this representation to a remote viewer. Figure 5.1 shows the system running with four smartphones' live video streams.

To investigate the effect of videos in panoramic contexts on users' performance, we conducted a user study comparing our system with fully-panoramic video and conventional webcam video conferencing for two spatial reasoning tasks. Linking back to the initial hypothesis presented in Chapter 1, the aim of the study was to understand whether partially dynamic panoramic representation, such as the one presented by PanoInserts, can help user improve spatial understanding of remote places (i.e. H2 and

Figure 5.1: *A typical PanoInserts session. Two cameras, pointing at two users, are tracked using image features. Another two cameras, pointing at a white wall and a white-board, are tracked more crudely using a marker-based method.*

H4). Additionally, we were interested in confirming whether the proposed representation can be achieved quickly and easily, leveraging solely available hardware (i.e. H3).

While, to our knowledge, the literature that investigates the effect of videos in panoramic context is rather scarce, it is important to note that the CamBlend system by Norris *et al.* [NSQ12] presents a framework, and investigation, similar to the one developed for PanoInserts. However, in contrast to our system, CamBlend only employs a wide-angle FoV image (i.e. 180° degrees) as the context. In addition, the context gets blurred when a high-resolution focus window is dragged around it to reveal parts of the remote scene. In our study we could have compared our system against CamBlend; however, our interest lies in investigating aspects which are intrinsics of the visual representation, rather than in comparing our system with existing frameworks. To this aim, we decided to compare our system with webcam and panoramic video, which, theoretically, display less and more spatial information, respectively. These two systems represent the extrema of a teleconferencing continuum in which the highly portable, but scarcely immersive webcam based video-chats represent the lower end of the interval, while the highly immersive, but scarcely portable fully-panoramic systems represent its end point. With respect to this continuum, we were interested in assessing weather our representation could position itself in the middle, ideally joining the best aspects of both ends.

The remainder of this chapter is structured as follows. The motivation behind the study, and consequently the system, are introduced in the next section. The chapter continues with technical implementation details of our system, including camera tracking, image registration, and rendering. We will then present a user study addressing the fundamental implications for spatial perception over three video display modes – webcam, fully-panoramic, and our system – showing that PanoInserts provides a good compromise in terms of both spatiality and accessibility between expensive fully-panoramic video and conventional webcam conferencing. Finally, we will discuss implications and design considerations for varying spatial forms of video conferencing, exploring how they are perceived and how they influence users when performing spatial reasoning tasks. A video showing the system in action, as well as additional informational material, can be found on the system's webpage[1]. Please note that some of the images reproduced in this chapter are adapted from the author's own work [PSW+13].

[1] http://www.cs.ucl.ac.uk/research/vr/Projects/PanoInserts/

5.1 Motivation

The quality and pervasiveness of cameras on mobile devices continues to increase. Most new laptops have a built-in camera, and most new smartphones and tablet-style devices have both front- and rear-mounted cameras. Rear-mounted cameras on mobile devices aim to replace or supplement the use of a point-and-shoot camera, while front-mounted and laptop cameras are often used for face-to-face video conferencing.

To this end, mobile devices have enabled portable video teleconferencing. Due to the portable nature of the devices, users may move around their environment and reposition cameras freely. In contrast, highly-developed video conferencing systems such as Cisco TelePresence [Cis06] are designed to support group collaboration, and feature multiple cameras and displays to achieve gaze awareness and a sense of space. However, such systems require equipment to be installed in a dedicated meeting room and also impose constraints on where participants position themselves to maintain gaze awareness during communication [Che02]. Panoramic video conferencing, as discussed in Section 4.2 and [RGC01], uses omnidirectional cameras such as the PointGrey Research LadyBug3 to capture a surrounding representation of a remote space and the people within.

The high-end systems described above are both expensive and lack portability, while the ubiquitous webcam-style video chat cannot easily transmit spatial relationships between several people or objects due to cameras typically having narrow fields of view. To overcome these limitations, we developed a system that we call PanoInserts. The system aims to support portable spatial video conferencing that lies between these two approaches in terms of both spatiality and accessibility. We aim to support meetings and other small-group interactions using only common personal devices communicating over the Internet. The system captures and transmits the visual representation of a real-world location and the people within for display to a remote viewer. It takes advantage of the pervasiveness of smartphones to create hybrid surround video communication in which a static panorama is augmented with live video inserts. As our system uses readily-available personal mobile devices, it can be rapidly configured and initiated, and lends itself to ad-hoc and spontaneous telecollaboration scenarios, such as the ones envisioned by BEAMING.

PanoInserts can be classified as a focus+context system: a system that shows a subset of information in full detail within a wider context of surrounding lower-density detail [BGBS02]. As such, PanoInserts presents a novel way to link together several videos to support remote meetings. Ideally our representation is able to convey more information than conventional web-cam style video chat, and the same dynamics that are encoded in a fully panoramic video. Nevertheless, the advantages of its representation are not immediately clear, and, to the best of our knowledge, have been studied only in an handful of prior works, the most notable being CamBlend by Norris *et al.* [NSQ12]. Therefore, we decided to run a study with the aim to understand whether partially dynamic panoramic representation, such as the one presented by our system, can help user improve spatial understanding of remote places. Specifically, we are interested in understanding whether PanoInserts can support fully panoramic spatiality while maintaining web-cam style video chat accessibility. To do so, we designed two tasks which involved spatial

reasoning, and we studied how users' performance varied across different video modes that included our novel representation, a fully-panoramic video and conventional web-cam video.

5.2 Architecture Overview

To outline the system's usage, and facilitate the reader in the rest of this chapter, we will now introduce a typical usage scenario for PanoInserts. Before outlining the scenario, it is important to understand what are the critical aspects of communication which are seen as being important to cooperative work. In their seminal work on spatiality and "shared spaces" [BBRG96], Steve Benford and his colleagues identify a range of issues which are critical for successful remote meetings. These include: the importance of creating explicit, familiar and persistent environments within which cooperative work can be situated; The importance that participants can establish a general awareness of what others are doing beyond their current focused activity [HL91, HRS92]; The importance to exploit people's natural understanding of the physical world, including spatial factors in perception and navigation, in order to construct cooperative systems which can be more easily learned and used [CBMW91]; And the importance of establishing clear and common shared spatial references through which situating the collaboration. The following scenario then, takes these aspects into consideration, and shows how we designed our system to support them.

Imagine a typical video-conferencing session in which a group of people (i.e. the *locals*, borrowing from BEAMING's terminology) in one city would like to have a technical discussion with a colleague located in another city (the *visitor*). This scenario is one of the typical use case for BEAMING, and it has already been introduced in Section 4.2.3. In the minutes prior to the conferencing session, one of the locals captures a panorama of the meeting room using built-in software on their smartphone. Subsequently, each local places their own smartphone in front of them so that its front camera points towards their seated position and the rear camera points at a marker (see Figure 5.4(a)). The visitor receives the live video streams from all locals' smartphones registered on the captured panorama. The visitor receives a surrounding representation of the meeting space, and hence can see the locals' seating arrangement and where each person is looking. During the discussion, the visitor asks the locals to draw a diagram to clarify some technical details. One of the locals repositions her phone to point at a white-board located in the meeting room and walks over to draw the diagram. The video-feed from the moving smartphone camera is tracked and re-registered within the panorama to present a live view of the white-board. Meanwhile, one of the still-seated local explains the diagram. The visitor can see both points of interest in the transmitted panoramic representation of the room, and can simultaneously interact interact with both.

Following BEAMING's main principles, our system design is motivated by the goals of accessibility and practicality. The system should be accessible in the sense that a meeting place should not require cumbersome tracking equipment, cameras, or dedicated networks. Rather, the required hardware should be commonly available smartphones and computers connected to the Internet. The system should be practical, meaning that it should be configurable in less than five minutes and should be dynamically reconfigurable during use. This implies that users are able to connect, disconnect and reposition smart-

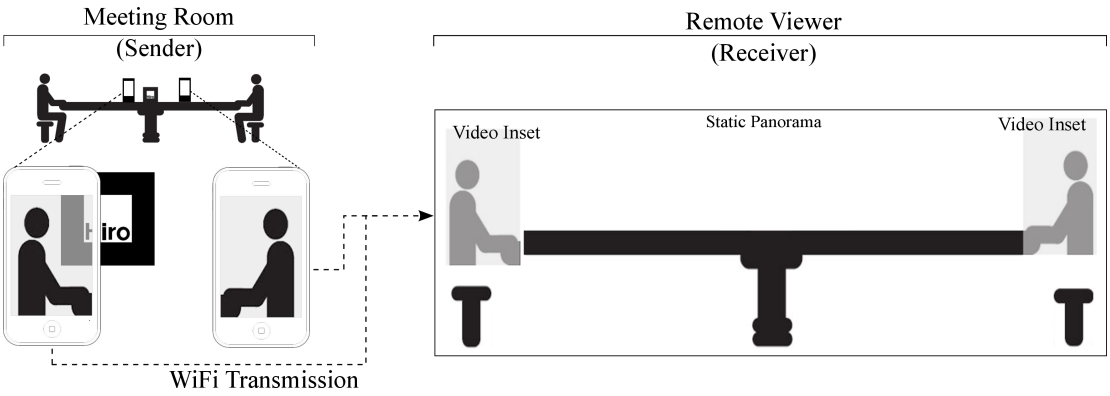


Figure 5.2: *Architecture overview. In the meeting room, the smartphone on the left is performing marker-based camera tracking and transmission of both camera pose and video, while the smartphone on the right is streaming only video. The remote viewer, which runs on a standard PC, receives this information and a) inserts a video stream based on the rough marker-based location (on the left) and b) performs feature-based camera tracking and accurately positions the corresponding video (on the right). Both videos are overlaid onto the previously captured static panorama of the meeting room.*

phones during the session. Allowing repositioning is particularly useful in situations where people are moving around the environment or when there are fewer available cameras than there are potential points of interest.

We use the video acquired from mobile phone cameras to transmit and dynamically insert views of the remote location within a static panorama. Our system comprises of three main modules: camera tracking, transmission and display (see Figure 5.2 for an overview of the system). The sender side features gross camera tracking based on marker (phone on the left in the figure) and transmission of both camera poses and video streams. The receiver side is responsible for computing an accurate feature-based camera tracking and receiving, integrating and displaying together multiple videos from multiple cameras. In addition to this, our system requires a preliminary stage for acquisition of panoramas. This additional step can be performed by using any desired software, including additional software running directly on the phone.

The software running on the smartphones (i.e. the sending side) was written using ARToolkit for iOS [ART03] and runs on devices running iOS4 or higher. The receiver-side software runs on PCs running Windows XP or higher, and uses the OpenFrameworks framework [Ope06], which uses OpenGL for rendering. Finally, for the feature-based camera tracking we employed OpenCV [Wil99] and the SiftGPU package [Wu07], a GPU implementation of the SIFT algorithm.

5.2.1 Construction of Panoramas

Many tools exist to assist in the construction of panoramas (see Section 2.3.1). While PanoInserts does not constrain the construction to any specific technique, it assumes that the panorama is available as a cube map, for display purposes. This, however, is not a limitation of the system, as conversion between panorama types can be easily performed. For the user study we run with the system, we used a cube-map with six faces each 2048×2048 in resolution (see Figure 5.3), assembled from 36 images using the PTGui software [New01]. However, the panorama could have been built also with software readily avail-

Figure 5.3: *Static cube-map panorama. Note the absence of furniture.*

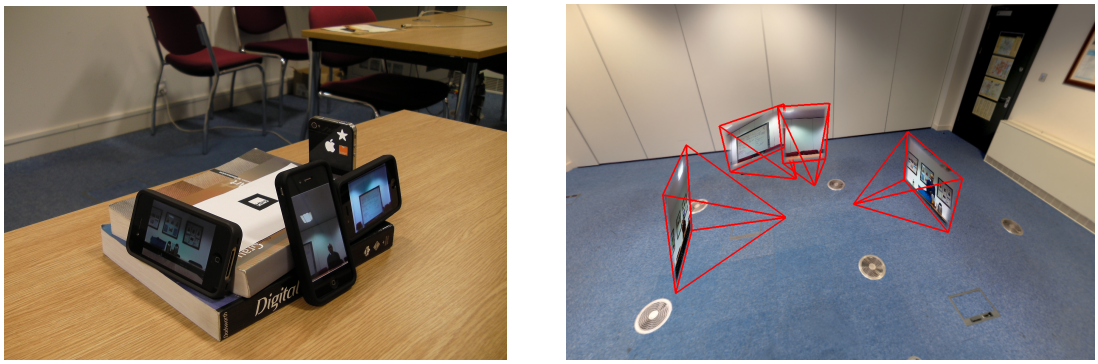
able on the phone, such as Microsoft's Photosynth [Mic08] or Android or iOS built-in image stitching applications.

5.2.2 Camera Tracking

The system relies on two tracking approaches to ensure that the camera frame is displayed correctly within the panorama. The system's preferred choice of tracking is a feature-based tracker that is run on the receiver. This approach is used when enough image features can be extracted from the video streams. The other approach is based on a single marker, and it is used during the system setup or when the more accurate feature-based tracker fails, such as featureless areas or in situation when the video quality is poor. Our system supports both automatic and manual selection of the tracking type. Users can either manually switch between tracking techniques by touching the screen, or have the system automatically choose the best tracking solution. If automatic selection is enabled, the system uses the device accelerometer to assess whether the unit is moving or not, tracking the marker only when the phone is static.

Marker-based Tracking

Ideally, we would like to track the cameras solely by registering the images captured against the panorama, as this would allow the users in the environment to have full control over the cameras. However, there are several barriers in doing this. First of all, our panoramas are only roughly accurate: furniture and other objects might move or the lighting might change. This is a common problem to any system that uses feature-based tracking. Second, our envisaged capture spaces (i.e. indoor scenes) often contain large feature-less areas (e.g., white walls in Figure 5.1) which would not be amenable to direct or feature-based image alignment methods. Third, our scenes contain moving humans and other objects



(a) *System setup: configuration of four smartphone cameras around a marker.*

(b) *3D positions of the cameras estimated from marker tracking. The marker is placed, roughly, in the center of the panorama which is drawn as the background*

Figure 5.4: *PanoInserts marker-based tracking.*

that move and change appearance (e.g the white board, which is on wheels, and the locals in Figure 5.1). In addition to this, we note that the quality of video available on mobile phones is usually low: under motion, the image is blurred and focusing and exposure balancing are slow.

Whilst some of these issues could be tackled by integrating other forms of camera tracking, such as built in inertial measurement unit (IMU) data as in Nyqvist and Gustafsson [NG13], this is not a robust option over long periods. Such solutions tend to accumulate large tracking error over time. Instead, we decided to employ a marker based camera tracking that computes a gross camera pose estimation. Such estimation is enough to initially display the video frames in their correct location, with a relatively small error, and can be obtained with negligible computational time (Figure 5.5(a)). We exploit the fact that recent phones, such as the iPhone 4, have two cameras. This allows us to stream the video to augment the panorama from the front (display-side) camera, and to track the marker using the rear-side camera. We decided to employ the front camera video for the streaming so that the users can see the video that is being transmitted while operating the device. Our system only requires a single marker in the environment, placed roughly in the center of the remote location (Figure 5.4). It is important to note that placing the marker roughly in the center of the remote location ensures that all the cameras that can see the marker roughly share an optical center. If the marker is also at the center of the panorama, then this guarantees that all the cameras will fit to the panorama.

Feature-based Tracking

Video registration based solely on marker-based tracking is only roughly accurate, resulting in a crude camera pose estimation. The next stage, then, is to refine such estimation by employing a more precise feature-based tracking algorithm (Figure 5.5(b)). This step effectively means registering the camera image to the relevant face(s) of the cube-map. The registration requires the estimation of a homography that maps the video frame into the face of the cube-map that has most overlap. To find this homography, we robustly estimate the features matching within two views employing SIFT features [Low04] and RANSAC refinement [FB81]. We opted for SIFT descriptors as they are invariant to different geometric transformations (scaling, rotation and translation) and they also provide a very robust match across a

(a) *Marker-based tracking.* **(b)** *Feature-based tracking.*

Figure 5.5: *Results from different camera tracking methods.*

large range of additional of noise and change in illumination.

When setting up the system, we pre-calculate and store SIFT descriptors for each of the six cube-map faces. As a new video image is received, from the last rough camera position given by the marker tracking we can filter out some of these SIFT descriptors from consideration to help removing false matches due to room symmetry and repeating elements. We do this crudely and conservatively by storing SIFT features in octants in the azimuthal plane, and only considering the two octants that most overlap the camera volume in that plane. We then extract the features from the received frame and calculate the number of matches of these features against the filtered sets for all six cube-map faces. We take the face with the largest number of matches and refine the corresponding matches using the RANSAC algorithm. See Figure 5.6 for an illustration of this process. Since RANSAC could excessively reduce the data set, we try to ensure a sufficient number of matches (eight – double the minimum number of points needed to evaluate any homography) by incrementing the acceptance error threshold in RANSAC until the criterion is met or the error threshold becomes too large. Finally, the parameters of the mapping homography H are evaluated from the robust point matching set using the gold standard algorithm [HZ04]. Because registration can fail in featureless areas, we check that the homography is reasonable (i.e., not degenerate or scaled by very small or large values). For videos where registration fails (e.g., due to insufficient matches or degenerate homography), we fall back to using the position given by the marker tracking.

5.2.3 Transmission

The transmission module is responsible to transmit marker-based camera poses and video streams, from the sender to the receiver. This information is not necessary streamed together, and a packet can contain camera pose only, video only, or a combination of the two. Transmission is performed over UDP. In the current implementation, video is read at 480×360 resolution, using JPEG encoding for each frame. This design decision was constrained by the fact that the operative system of the devices used, iOS4, neither gives direct access to the raw image data nor allows for different compression methods. Nevertheless, each video packet, sent at a rate of 10 Hz using a shared wireless 802.11g network, is typically 5–30KB, and thus within the capacity of a single UDP packet. On the receiving side, the system receives a number of input video sequences and corresponding estimates of the camera pose relative to the panorama. This

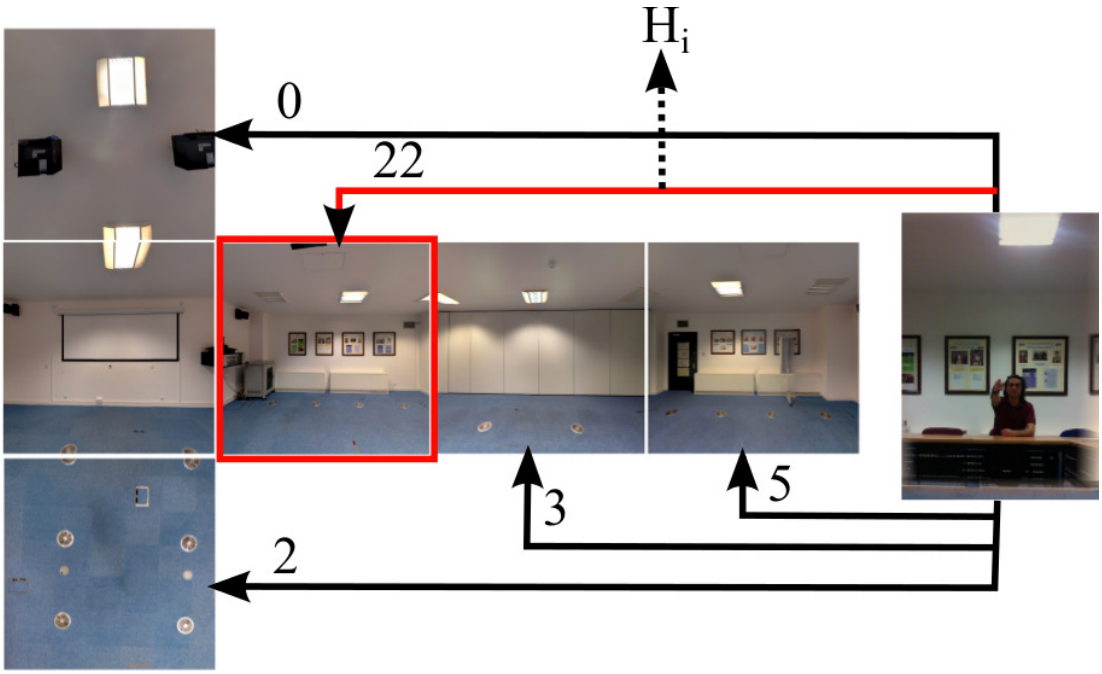


Figure 5.6: *Feature-based tracking. We extract a set of features from an incoming video frame and we match it against the pre-stored features of the cubemap's faces. Numbers in the figure (which are arbitrary and for illustration purpose) represent the matching results. We then take the face with the largest number of matches and refine the corresponding matches using the RANSAC algorithm. Finally, we estimate the homography H_i from the resulting matches using the gold standard algorithm [HZ04].*

information is then used by the receiver to correctly display the various video streams within the static panorama.

5.2.4 Display

The renderer integrates multiple videos from multiple cameras, displaying them in a 3D scene with the panoramic image as background (Figures 5.1 and 5.5). As the renderer operates on the information received from the sender, the rendering varies depending on the type of packet received and is computed for each camera separately.

If the received packet contains the marker-based estimate of the camera pose and a video frame, then the renderer displays the video inset using a projective texture based on the camera position returned by the marker tracking. The texture is projected on the six faces of the cube-map, and it is applied to a camera volume which is shaped by the intrinsic parameters of the smartphone's front camera (Figure 5.5(a)). If the receiver receives only a video frame, then the feature-based camera tracking needs to be performed to estimate the camera position. When this is done, the renderer applies the incoming video as texture of an extended plane that coincides with the face of the cube-map that is selected by the SIFT matching process. The estimated homography is converted into a texture coordinate matrix, and this plane is rendered with the video textured on it over the original texture from the static panorama. To obtain visually pleasant video overlay, the incoming video texture is blended into the panorama using alpha blending around the borders of the video texture. Furthermore, as the color balance of the smartphone's front camera might be noticeably different from the camera used to captured the panorama

images, we ensure the white balance is the same by computing beforehand an overall static color balance correction using example images (Figure 5.5(b)). We do this by sampling both the camera images and the static panorama on a 10 by 10 grid at matching points, and we then estimate the correction factor independently for each colour channel. The correction results in a constant scaling of the phone's image colour by a factor of (1.0, 0.97, 0.95).

5.3 User Study

Our user study aimed to assess the extent to which viewers are able to perceive and act on varying video modes over two spatial visualization tasks. Specifically, we are interested in evaluating the benefits of videos in panoramic contexts when compared to other video modes. To this aim, we compare our system with webcam and panoramic video, which, theoretically, display less and more spatial information, respectively. To be consistent with the webcam condition that features the usual single camera, we test our system with only a single smartphone. For both webcam and PanoInserts conditions, we used the iPhone 4 front-facing camera in portrait mode to capture and transmit video. While our system is able to support several smartphones running in parallel to populate a static panorama with dynamic inserts, it is critical to assess the quality of our fundamental approach without being diverted into assessing how this may change as the number of dynamic inserts increases. We used a PointGrey Research Ladybug3 camera for the panoramic condition (see Section 3.2.1 for hardware specifications). To be consistent with the PanoInserts condition, we used an equirectangular projection to render the video acquired from the Ladybug3.

Deciding how to adequately evaluate a novel system is non trivial, and requires a clear understanding of how the technology will be used. Only by clearly establishing which activities a technology is designed to support an adequate and appropriate evaluation of it can take place. Typically, this practice allows the developers to identify "critical parameters" that can then be evaluated, and through which system's performances can be tested [New97b, New97a]. Critical parameters, a concept that figures constantly in design literature [Rog83, Vin91], provide the designer with a primary unit of performance against which to predict or measure the system ability to meet a set target. They provide a measure in terms of the purpose of the system, rather than in terms of its functional design. Critical parameters therefore provide a direct and manageable measure of the system ability to serve its purpose.

Selecting truly representative critical parameters, and identifying the type of activities that a system is designed to support, can then facilitate the definition of the tasks employed for its evaluation, and ultimately will ensure the ecological validity of the study. In our case, we designed PanoInserts with the goal of supporting ad-hoc, shared meetings that require a good level of spatial understanding. Such type of meetings include remote assistance [RBB06, FKS00, BRB10], in which an "expert" instructs one of many users on how to accomplish certain tasks, as well as meetings that take place in remotely shared environments that feature multiple participants and additional tools [FGR04]. In the case of remote assistance, spatial references are usually used by the expert to guide the other person to fulfil certain tasks, such as picking tools, assembling objects or identifying correct areas to act on. Similarly, in the case of more general meetings, spatial references are often used to address other people that take

part in the conversation, or to direct the attention to tools which may be outside the current FoV, such as whiteboards or posters. There is also another class of meetings which could benefit from spatial-understanding. These are more casual meetings in which users want to visually share particular areas of their environments which are not captured by the current camera's FoV. For instance, image a person using our system to show his/her new office to his/her colleague. The colleague can ask the user to move the camera towards certain areas, using objects as references. If multiple phones are used, then the user can always be in view, while other phones are moved around to reveal different parts of the office.

Therefore, we argue that to correctly evaluate our system with respect to its real-world usage, tasks requiring a high level of spatial understanding should be employed. Additionally, these tasks should be evaluated trough critical parameters that can address the level of spatial thinking achieved by users. Such parameters should focus on the precision of users while interacting with the remote environment, for instance while manipulating objects present in the shared space, or on their ability to makes spatial-references to it.

With this goal in mind, prior to running the experiment as it is documented in this chapter, the experimental design was refined over several iterations in order to select adequate tasks for the evaluation. Initially, we listed a number of possible tasks to evaluate the effect of video modes on spatially-related tasks. These included arranging furnitures in a room, draw a map of a remote place and answer location-related questions about a room. However, during our brainstorming session it quickly emerged that, to investigate the operational benefit of using videos in context in space-focused scenarios, the experimental task must require the user to explore the remote space and interact with objects there to mimic natural interactions. Additionally, we concluded that the objects should be located in different areas of the remote location, so that they sample the entire space, to uniformly study the effect of different video stimuli.

Therefore we decided to split our study over two tasks. Both tasks involved object placement: either placing virtual objects to match the locations of real objects as perceived from the video stimuli, or the reverse of this, which is instructing a confederate to place real objects as seen through video stimuli to match the locations of virtual objects. Despite the somewhat artificial nature of the tasks, we feel they remain representative of both remote assistance and meeting scenarios, that involve referencing local objects and require a strong understanding of remote spaces. In Section 4.2.3 we introduced a discussion on spatial deixis and on how these affect, and facilitate, communication. This is particularly true for the type of scenarios supported by PanoInserts, in which being able to correctly localise, manipulate and reference objects and parts of the shared environment is paramount to achieve successful collaboration. Such considerations are also supported by previous HCI research, where there is a precedent for the use of tasks similar to the ones employed in our study [LYKH11, LHK+03, YCNB96, NSQ12, SJF+13].

Hypotheses. In both tasks, we measured object placement error, task completion time and, in two of the video modes (webcam and PanoInserts conditions), requested camera movements. Additionally, we collected the results of two post-experiment questionnaires, one focused on system usability, and one more focused on the tasks.

For both tasks we expected task performance to vary according to the spatial information each mode theoretically preserves. Hence, we expected participants using the panoramic video to be able to both place objects (virtual object placement task) and instruct objects to be placed (real object placement task) more accurately than participants using PanoInserts. In turn, we expected participants using PanoInserts to be more accurate than those in the webcam condition. Regarding number of camera movements, we expected the participants using PanoInserts to require fewer than those in the webcam condition due to the presence of the static panorama background, which in theory should enrich the spatial information conveyed by the system. Note that the panoramic condition requires zero camera moves as the whole panorama is dynamic. Regarding task completion time, we expected that participants using the panoramic video would require the least time than those in the other two conditions. Our expectancy of the usability scores as measured by one of the questionnaires were less clear, as the panoramic representations of space as presented by both PanoInserts and the panoramic systems may be unfamiliar to participants and take some acclimatization that may influence the scores. We did expect, however, that all three video modes would be ranked reasonably highly in terms of overall usability. Finally, we had no clear expectations on the task-related questionnaire, as we believe this is directly related to task performance's perception, which however may be influenced by the unfamiliarity of the panoramic representation.

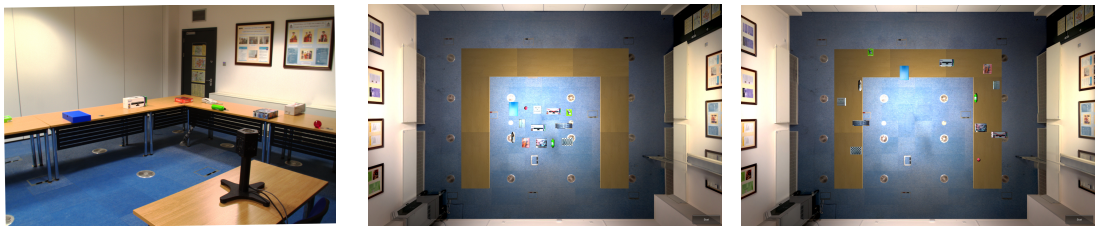
5.3.1 Method

Participants

The study involved three video conditions, and a total of 36 unpaid participants took part to it (12 in each video mode). We alternated the order in which the two tasks were performed to minimize the influence of learning effects, and we randomly assigned each participant to video mode. Participants performed both experimental tasks in a single video mode, so the experiment featured a between-subjects design in terms of the independent condition of video mode, and a within-subjects design in terms of task. Participants were recruited from the staff and student population at our university, via e-mails.

Design

The tasks adopted in the study intended to explore the accuracy with which participants can correctly obtain a spatial understanding of a remote environment over the three modes. Both tasks involved object placement: either placing virtual objects to match the locations of real objects as perceived from the video stimuli, or the reverse of this, which is instructing a confederate to place real objects as seen through video stimuli to match the locations of virtual objects. Hence, in both tasks, we measured object placement error, task completion time and, in two of the video modes (webcam and PanoInserts conditions), requested camera movements. After the participant had finished each task, we measured the positional (2D horizontal) error of either the virtual objects as placed by the participant in the virtual room (first task), or the real objects as placed by the confederate as per the participant's instructions in the real room (second task). Following the experiment, participants completed the standard System Usability Scale (SUS) questionnaire, which gathered subjective assessments of usability of the three systems; for



(a) *The real environment.* **(b)** *The VE with empty tables.* **(c)** *The VE with object on the tables.*

Figure 5.7: *Real environment and virtual copy used for the experiment.*

the full set of questions, please refer to Brooke [Bro96] or Appendix D. Additionally, participants were asked to answer five task-related questions (see Appendix D for a list of questions), and their impression on the system was also recorded.

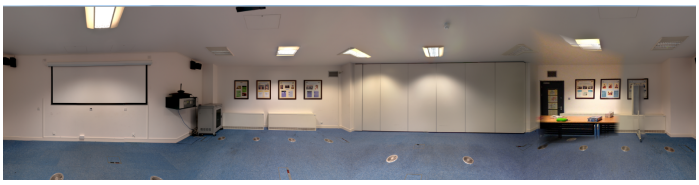
Procedure

As already mentioned, our study was split over two tasks. In both tasks, the participants viewed a remote meeting room featuring a "horseshoe-shaped" table arrangement surrounding a central table on which the appropriate camera could be positioned (Figure 5.7(a)). We used stands to ensure that video from the Ladybug3 or iPhone camera was acquired from the same position. All the cameras were initially facing the center of the room. The set of objects, for a total of thirteen objects, consisted of typical things one may find in an office or at home, such as water bottles, phones or boxes, and varied in size from 10 cm^3 – 50 cm^3, and in color and shape.

The first task required participants to view a remote meeting room in which thirteen objects were positioned on tables around the room. Participants were required to determine where these objects were positioned in the room, and to use an interactive virtual model of the room to position the objects' virtual counterparts accordingly. A scaled virtual model of the room was created using Autodesk 3DS Max, which was then loaded into the experimental interface developed using Unity [Uni05]. At the beginning of the experiment, the virtual objects were located at the center of the virtual model shown in Figure 5.7(b). The virtual objects could be repositioned by dragging-and-dropping using the mouse. As the angular separation between the leftmost and rightmost objects was approximately 180°, participants in both the webcam and PanoInserts modes required the 30° camera to be rotated during the task to reveal different areas of the room. Hence, in these two conditions, participants could instruct a confederate located at the remote meeting room to rotate the camera.

The second task reversed the real-to-virtual object placement done in the first task, and required participants to match the positions of real objects in the meeting room with those presented in the same virtual model as used in the first task. Participants viewed a non-interactive virtual model of the remote meeting room in which the same thirteen objects were positioned (differently to how they were positioned during the other task) as shown in Figure 5.7(c). Participants instructed a confederate at the meeting room to place objects to match the virtual layout. The objects were all placed in the middle of the real room. However, the participants did not have to locate the objects first, but rather they only had to ask the confederate to pick a specific object to start its positioning. To minimize the influence of the

(a) *Panoramic.*



(b) *PanoInserts.*

(c) *Webcam.*

Figure 5.8: *Representations of the remote room using each system. For both the Panoramic and PanoInserts condition an equirectangular projection is used for rendering the panoramic imagery. Please note that during the experiment the chairs visible in (a) were removed from the environment.*

confederate's behaviour, they could only follow direct instruction from the participant such as, "place the object X half-way along the table directly behind you", and could not help in any other way. The confederate strictly and literally followed such directions given by the participant with minimal verbal interaction. As in the first task, participants could also ask the confederate to rotate the camera in the webcam and PanoInserts modes to reveal different parts of the scene.

The room used in our experiment is a popular meeting room at our university, and therefore some participants had previously been in it. However, others had never been into the room before, and therefore the level of prior knowledge of the space varied across the population. Hence, to ensure that all participants had similar prior knowledge of the remote environment, before the experiment we gave each as much time as they liked to walk around the room (cleared of all objects) and become acquainted with the space. This ensured that no subgroups within the population held more information about the room than the rest of the participants, allowing for a fair comparison of their tasks result.

The participant was then brought into the lab where he/she was presented with two workstations: one displaying the video-mediated representation of the room in one of the three video modes (Figure 5.8), and the other displaying the virtual representation of the room. Objects were arranged in both real and virtual environments to the appropriate starting arrangement depending on which task was to be performed first. The participant was briefed on the appropriate task and on how he/she may instruct the confederate to move the camera in the webcam and PanoInserts condition and also to pick up and place objects if they were performing the real object placement task. Following completion of the task, the object placement errors along with time taken and number of camera moves (in webcam and PanoInserts conditions) were recorded. The room was then rearranged for the remaining task. The participant was briefed on the remaining task which they would then carry out, and data recording was subsequently performed. Finally the participant completed the questionnaires and his/her impressions on the system, if any, were recorded.

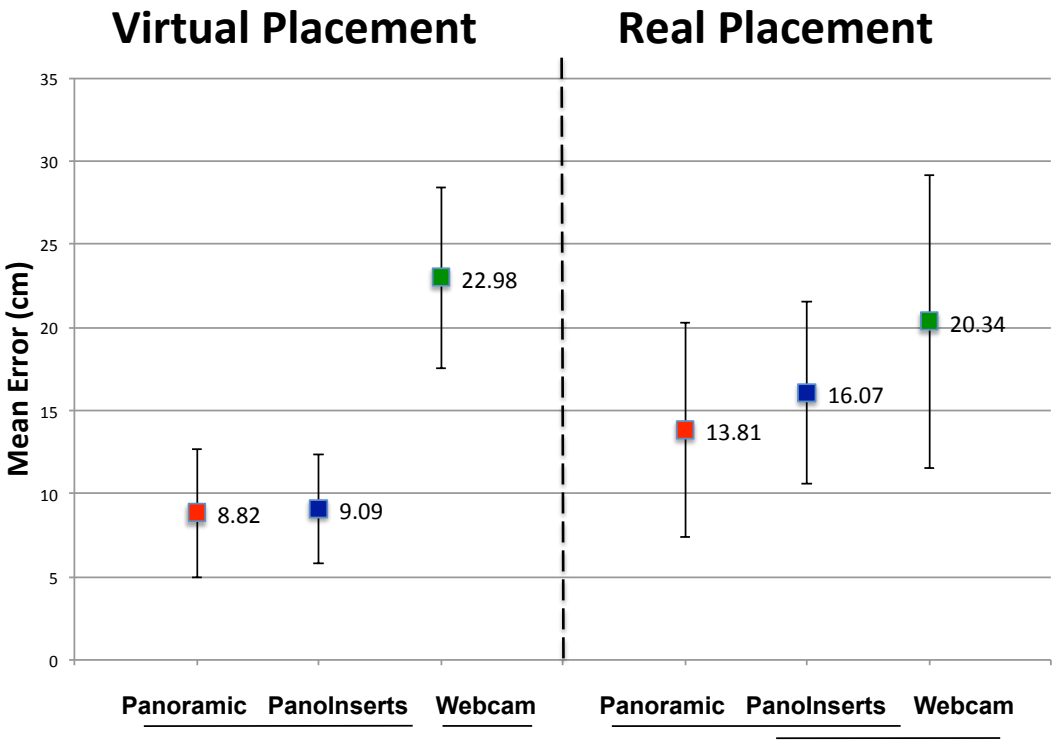


Figure 5.9: *Mean object placement error and standard deviation for the three systems in both tasks. Conditions jointly underlined are statistically similar.*

5.4 User Study Results

The primary dependent measures of interest used for both tasks were the accuracy expressed as errors in object placement, the time taken to complete the task and, for PanoInserts and webcam conditions only, the number of camera moves requested to complete the task. Initially, for statistical analysis a 3×2 (video $\times$ task) mixed Analysis of Variance (ANOVA) was computed using SPSS [IBM09] to analyse each of the dependent variables. Video mode was a between-subject factor, while task was a within-subject factor. A significance level of .05 ($\alpha = 0.05$) was used for judging the significance of effects. No samples were removed from our analysis, as all the participants successfully completed their tasks.

5.4.1 Placement Accuracy

Accuracy shall be considered first. Figure 5.9 shows the mean error and standard deviation of object placement error for both tasks. We first address the task in which participants were required to place objects in the virtual environment to match the real environment's arrangement while viewing the meeting room using one of the three video modes (we shall call this task the *virtual object placement* task hereafter). We observed a lower error for the panoramic ($M = 8.82$ *cm*, $SD = 3.86$ *cm*) and PanoInserts ($M = 9.09$ *cm*, $SD = 3.30$ *cm*) conditions than for the webcam condition ($M = 22.98$ *cm*, $SD = 5.44$ *cm*). We now focus on the task in which participants were required to instruct a confederate to place objects in the real environment to match the virtual environment's arrangement while viewing the meeting room using one of the three video modes (we shall call this task the *real object placement* task hereafter). As before, we found a lower error for the panoramic ($M = 13.81$ *cm*, $SD = 6.46$ *cm*) and PanoInserts

($M = 16.07$ *cm*, $SD = 5.47$ *cm*) conditions than for the conventional webcam condition ($M = 20.34$ *cm*, $SD = 8.80$ *cm*).

To further analyse the dependent variable of accuracy, we computed a 3×2 (video × task) mixed ANOVA using SPSS. Video mode was a between-subject factor, while task was a within-subject factor. Results showed both a main effect of video type ($F_{(2,33)} = 34.811$, $p < 0.001$) and task ($F_{(1,33)} = 9.725$, $p = 0.002$) on accuracy. Similarly, the interaction between video mode and task was significant ($F_{(2,33)} = 8.829$, $p < 0.001$). Simple follow-up main effects analysis showed that users in the panoramic condition were significantly more accurate in the *virtual object placement* task ($p = 0.002$) than in the *real object placement* task. The same emerged for the PanoInserts case ($p < 0.001$), but not for the conventional webcam conditions ($p = 0.110$);

Finally, to break down the effect of video mode at each level of task, we calculated two ANOVAs (one per task) using SPSS with the two factors of video mode and object and the dependent variable of placement error. Regarding the *virtual object placement* task, a significant main effect of video mode was found ($F_{(2,33)} = 66.555$, $p < 0.001$). Post-hoc Tukey tests revealed non-significant differences between the panoramic and PanoInserts conditions ($p = 0.979$), and significant differences between the webcam and panoramic conditions ($p < 0.001$). A main effect was found between PanoInserts and webcam conditions ($p < 0.001$). Additionally, a significant main effect of object was found ($F_{(12,33)} = 3.015$, $p < 0.001$). Moving to the *real object placement* task, a significant main effect of video mode was also found ($F_{(2,33)} = 4.849$, $p = 0.008$). Post-hoc Tukey tests again revealed non-significant differences between the panoramic and PanoInserts conditions ($p = 0.555$), and significant differences between the webcam and panoramic conditions ($p = 0.007$). However, no main effect was found between PanoInserts and webcam conditions ($p = 0.112$). The main effect of object was also significant ($F_{(12,33)} = 3.022$, $p = 0.001$).

5.4.2 Time to Complete

We will now focus on the time to complete the tasks. Figure 5.10 reports the mean time to complete, and standard deviation, for each task in each video mode. Regarding the *virtual object placement* task, participants were faster in completing their task when in the conventional webcam ($M = 169.92$ *sec.*, $SD = 52.63$ *sec.*) and panoramic conditions ($M = 198.37$ *sec.*, $SD = 60.81$ *sec.*) than in the PanoInserts condition ($M = 395.19$ *sec.*, $SD = 136.55$ *sec.*). As for the *real object placement task*, participants were faster in completing their task when in the panoramic ($M = 444.55$ *sec.*, $SD = 123.21$ *sec.*) and PanoInserts conditions ($M = 538.32$ *sec.*, $SD = 180.14$ *sec.*) than in the conventional webcam condition ($M = 561.01$ *sec.*, $SD = 237.97$ *sec.*).

To further analyse the dependent variable of time to complete, we computed a 3×2 (video × task) mixed ANOVA using SPSS. Video mode was a between-subject factor, while task was a within-subject factor. Results of statistical analysis found both a main effect of video type on time to complete ($F_{(2,33)} = 5.236$, $p = 0.011$), and a main effect of task on time to complete ($F_{(1,33)} = 72.079$, $p < 0.001$). Similarly, the interaction between video mode and task was also significant ($F_{(2,33)} = 6.001$, $p = 0.006$). Simple main effects analysis showed that users in the panoramic condition were significantly faster when

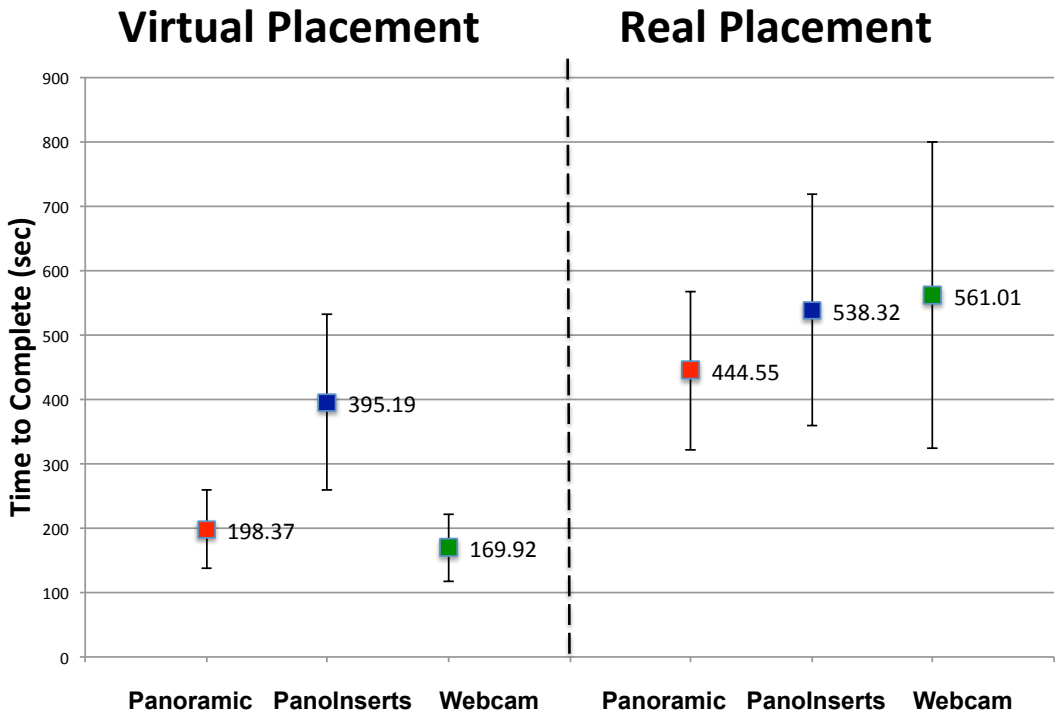


Figure 5.10: *Mean completion time and standard deviation for the three systems in both tasks.*

	Camera Moves	
	VOP	ROP
Panoramic	N/A	N/A
PanoInserts	7.58	7.41
Webcam	8.5	8

Table 5.1: *Mean required camera moves for the three systems in both the virtual object placement (VOP) and real object placement (ROP) task.*

performing the *virtual object placement* task ($p < 0.001$). The same emerged also for the PanoInserts ($p = 0.009$) and conventional webcam conditions ($p < 0.001$);

Finally, to break down the effect of video mode at each level of task, we computed two ANOVAs (one per task) using SPSS with the single factor of video mode and the dependent variable of total time to complete the task. For both conditions, video mode was not found to be a significant factor (*virtual object placement* – $F_{(2,33)} = 1.356$, $p = 0.272$; *real object placement* $F_{(2,33)} = 1.794$, $p = 0.190$). We note that there is a large variance between participants, and that we briefed participants to complete the tasks with object placement accuracy in mind as opposed to speed.

5.4.3 Required Camera Moves

For the PanoInserts and webcam conditions we also collected the total number of camera moves required by each participant while completing the two tasks. Table 5.1 reports the mean number of camera moves for each mode. Regarding the virtual object placement task, participants required less camera moves for the PanoInserts condition ($M = 7.58$, $SD = 1.62$) than for the conventional webcam condition

($M = 8.5$, $SD = 2.81$). Similarly, during the real object placement task participants requested less camera moves while using PanoInserts ($M = 7.41$, $SD = 2.06$) than while using the webcam video ($M = 8.00$, $SD = 1.80$).

To further analyse the dependent variable of requested camera moves, we computed a 2×2 (video × task) mixed ANOVA using SPSS. Video mode was a between-subject factor, while task was a within-subject factor. A significance level of .05 ($\alpha = 0.05$) was used for judging the significance of effects. Results of statistical analysis found no main effect of video type on camera moves ($F_{(1,22)} = 0.823$, $p = 0.374$), as well as no main effect of task on camera moves ($F_{(1,22)} = 1.600$, $p = 0.219$). Similarly, the interaction between video mode and task was also not significant ($F_{(1,22)} = 0.400$, $p = 0.534$). For both tasks, we also calculated an ANOVA with the single factor of video mode and the dependent variable of number of camera moves requested by the participant to complete the task. Regarding the virtual object placement task, no main effect was found ($F_{(1,22)} = 0.957$, $p = 0.339$). Focusing on the real object placement task, the ANOVA also did not uncover a significant difference between conditions ($F_{(1,22)} = 0.542$, $p = 0.470$).

Finally, for both webcam and PanoInserts conditions we computed the Pearson correlation coefficient r between the participants' requested camera moves and the participants' mean error. In doing so, we were interested in revealing the strength of the linear association between the two variables, to reveal whether to more camera moves would correspond higher accuracy. A moderate negative correlation was found for PanoInserts in both the virtual object placement task ($r = -0.664$) and the real object placement task ($r = -0.324$). However, for the webcam condition the correlation coefficient reveals a weak positive correlation for the virtual object placement task ($r = 0.126$) and a weak negative correlation for the real object placement task ($r = -0.104$). Implications of these results are discussed in the next section.

5.4.4 Questionnaires

Following the experiment, each participant completed the SUS questionnaire. All modes obtained positive results, with the webcam condition obtaining the best score ($SUS = 82.5$), followed by the panoramic ($SUS = 77.29$) and PanoInserts ($SUS = 73.54$) conditions. Based on these results, and following the analysis technique suggested in [LS09], the webcam system can be classified as Rank A system (out of six possible letter-grade ranks varying from A to F), while both PanoInserts and the panoramic mode can be classified as Rank B systems.

Regarding the task-related questionnaires, non significant differences emerged within video conditions or individual questions. Generally, all three modes scored similarly positive results, with a low standard deviation between average scores ($STDDEV = 0.27$).

5.4.5 Participants Comments

Following the experiments, we recorded participants impressions. Regarding the panoramic conditions, only positive remarks were registered. One participant thought it was hard to estimate the depth of objects located far away, and suggested he/she was using *"the wide field of view in combination with markers in the room such as air vents, tables and corners to align objects"*.

Impressions on the webcam conditions were more negative. A general remark was on the limited description of the whole environment. In particular one participant reported that *"my initial confusion was due to not knowing in which direction the camera pointed at the start"*. Another user commented that *"only being able to see a small section of the room at one time made it harder to estimate the position of objects on the tables"*.

PanoInserts' impressions were generally positive, with users consistently considering the panoramic background as a valuable tool to perform the tasks. Two particular informative comments reported that *"[...] by comparing the locations of features on the tables and walls it was fairly easy to judge the rough positions of the objects"* and that *"perspective can be a little bit confusing but the permanent items around (e.g., pictures on the wall) the room help to understand better the environment"*.

5.5 Discussion

5.5.1 Task Performance

The results from our user study reveal insights into the way participants were able to spatially perceive and act on information presented in the varying video modes. In both tasks, panoramic video and PanoInserts enabled greater accuracy than webcam video when positioning objects. This finding is in accord with each video mode's relative degree of spatiality as hypothesized, and suggests that both fully- and partially-dynamic panoramic representations of space can encode information that people can intuitively understand and act upon.

Exploring the number of camera moves participants performed reveals information about how participants went about completing the tasks. As the panoramic condition did not require camera movement, here we discuss only the webcam and PanoInserts conditions. While not found statistically significant in our analysis, participants in the PanoInserts condition performed fewer camera movements than those in the webcam condition (Table 5.1). A moderate negative correlation between camera moves and mean error was also noted for PanoInserts, but not for the webcam mode. This indicates that PanoInserts users were able to incrementally decrease placement error through camera repositioning. The same does not apply to the 2D video case, as its correlation coefficients reveal a weak positive correlation for the virtual object placement task. This suggests that participants could apply the additional spatial information presented in PanoInserts to improve their spatial reasoning ability of the remote location. Concerning the time to complete the tasks, PanoInserts' users systematically required more time to ultimate their tasks. This can be justified by the fact that the system performances was influenced by switching the camera tracking mode, which we will refine in future versions of the system.

Placement accuracy differed in between the two tasks, with the virtual object placement task resulting in a relatively smaller error and standard deviation than the real object placement task. While the two tasks were complementary and both relied on spatial reasoning, they differed in some key aspects. When positioning virtual objects to match those viewed in the physical space, participants observed a visual representation of the real objects spread over the tables in the room from a perspective similar to being in the room. This embedded additional spatial cues in the video stimuli, provided by the objects' relative

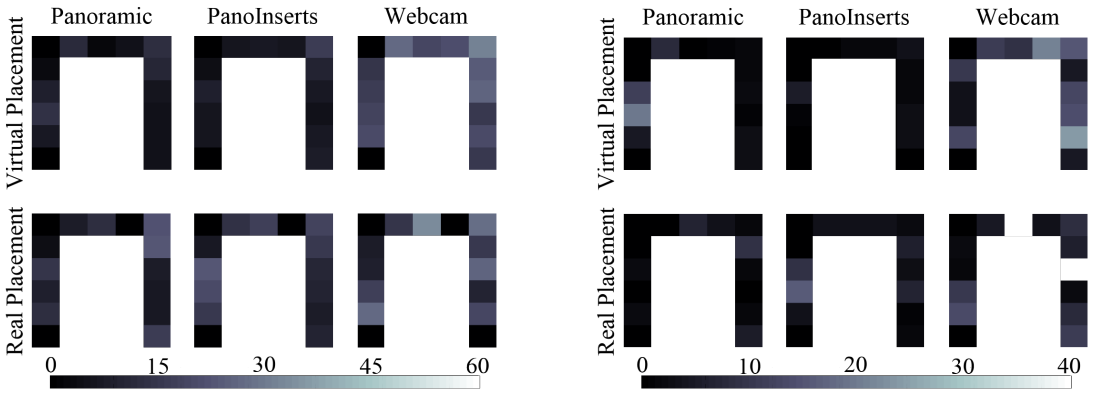
locations and the camera's viewpoint. This resulted in some participants instructing the confederate to move the camera "in between" certain objects, effectively restricting placement error to greater extent than in the real object placement task. Contrastingly, in the task requiring positioning of real objects to match those in the virtual space, participants were presented with a top-down virtual reference representation from which to work from that was more similar to the perspective of a CCTV camera than it is to being in the room. So, participants could use only environmental cues to estimate where an object should be placed. They could also use objects that they had just placed, but error could accumulate. This allowed more room for incorrect placement.

Hence, the two tasks presented qualitatively different reference stimuli from which the task of positioning objects is then required to be carried out. The accuracy results shown in Figure 5.9, and the results of the statistical analysis, show that participants found the real object placement task more difficult than the virtual object placement. Exploring the impact of task further, we calculated three post-analysis single-factor ANOVAs using task as factor, and data from a single video mode. Significant differences were found between tasks in panoramic ($p = 0.028$) and PanoInserts ($p = 0.001$) conditions, but not in the webcam condition ($p = 0.607$), where the real object placement task actually attained slightly greater accuracy. We note, then, that participants found the conversion between a person-perspective view to a top-down representation (as in the virtual object placement task) easier than they found the reverse. However this depends on the spatial richness of the stimuli, and does not hold if the spatial nature of the perspective view is impoverished as in the webcam condition. We now further explore the differing spatial representations offered by the three video modes.

5.5.2 Spatial Representation

When displayed on a standard flat display, panoramas represent a surrounding environment in a way that is often not intuitively clear, and differs considerably from how we visually perceive space in normal life. Panoramas present space at a greater FoV than the human visual system does, so the viewer has to cognitively translate that representation before understanding it. On the contrary, conventional webcam video presents space with a FoV that is less than human vision, so is directly intuitive for the viewer. While our experimental results show that people can understand the panoramic content and use it to complete the tasks efficiently, there are likely to be better ways of presenting it. In the following chapters we will explore both hardware and software approaches to this problem. Displays such as Global Imagination's spherical Magic Planet [Glo06], portable tablet devices or immersive projection technologies such as head-mounted displays will be investigated. Such displays types are able to complement the acquisition technology and present panoramic content in a way that preserves its surrounding nature. In combination with this, also software approaches to enable clearer representation of the spatial mapping between panorama and environment will be investigated in later chapters. In particular this can be achieved through visually-correcting interesting portions of the panorama through a "pop-out" metaphor, or by presenting the entire panorama in a virtual environment, as seen in [MSD+12].

As stated previously, participants visited the experimental meeting room prior to the experiment, and were also presented with the virtual model during experiment, helping them to form an idea of the



(a) *Mean error and room location.* **(b)** *Mean error variance and room location.*

Figure 5.11: *How mean error and error variance varies over the room. Each tile represents a portion of the desk.*

spatial layout of the room. During the experiment, participants were required to translate between a top-down virtual model of the room and a first-person perspective video representation of the room. These two visualizations present space differently. Specifically, the distortion present in the video modes varies across the image, so that the screen-space distance between two pixels in the video that map to two points in the physical room may not be equal to the distance of another two other points in the room of equal physical distance. This depends on the distance of the objects to the camera, and is due to camera foreshortening, which usually results in more error around the corners of a camera view.

We assessed the influence of object position post-hoc, and present Figure 5.11. The plots visualize the horseshoe-shaped table in the experimental room, and encode mean object placement error and error variance as a heat-map. Both error and error variance is seen to vary across the environment, with the greatest readings localized around upper-right corner and left side of the tables. The varying visual distortion inherent in video is likely to influence object placement accuracy around the 180° range. The error variance across objects (Figure 5.11(a)) is noticeably larger for the webcam condition than the other two conditions, suggesting that participants using it were performing the spatial reasoning task based on poorer information and were less accurate as a result.

5.5.3 Usability

All the three systems obtained a high SUS score, with participants rating the webcam mode highest ($SUS = 82.5$, Rank A), followed by panoramic ($SUS = 77.29$, Rank B) and PanoInserts ($SUS = 73.54$, Rank B) modes. The webcam system's higher score is likely due to its familiarity with participants. Also regarding usability, it was interesting to observe how participants went about the tasks in each condition. Participants in the webcam condition often required an initial camera rotation from one corner to the room to the other, indicating that they were unsure as to where the camera was facing in the room. Additionally, several participants in the webcam condition became confused with regards to which direction they needed to rotate the camera to see a different part of the room, which may indicate difficulty in self-localization in the remote location. These observations are supported by some of the post-experimental comments recorded. The majority of participants that experienced PanoInserts con-

sidered the static panorama to be a valuable resource providing spatial information about camera heading and object location.

5.5.4 Conclusion and Limitations

The experimental work described in this chapter reveals interesting insights on the quality and usefulness of videos in panoramic contexts for remote collaboration. In particular, results showed that augmenting a static panorama with live video insets can greatly improve on standard webcam videos when performing spatially-localised tasks. Results also showed that the proposed representation performs similarly to fully panoramic video, a video mode that represents the current state of the art for videoconferencing, albeit high prices and restricted portability.

Therefore, the outcome of the user study here presented can help us address one of the main questions that motivates the research. Videos in context can be considered as a valuable tool to enable remote spaces exploration and remote collaboration. PanoInserts showed that spatially localised video can be used to increase the spatial information transmitted during VMC, improving the quality of communication between users, but also enhancing their spatial thinking. In particular, using panoramas as a context can be considered as a special case of the general problem of aligning content to world model - a fundamental problem already faced in BEAMING for environmental reconstruction. By offering partially dynamic surrounding representation of a place, videos in panoramic contexts can greatly reduce the cognitive load required by users to perform spatial thinking. This, in accordance with the initial hypothesis (i.e. H2 and H4), means that users are able to understand and act upon the spatial information encoded within the proposed representation. Interestingly, the benefits of our representations come with little technical effort achievable with common devices, confirming another point of the initial hypothesis (i.e. H3). Panoramas can be acquired and rendered in a variety of ways, while registration of videos within the context can be performed at interactive rates on a variety of devices, including portable ones. Therefore, we can conclude that there is indeed benefit in using videos in panoramic contexts for telecommunication, especially in a multi-party interaction where dynamics might be spread over a large environment and cannot be easily captured by standard webcams.

While the work described in this chapters helps us addressing one of the main questions that motivates the research, it is important to note that some of the aspects which shaped the development and experimental investigation could have been carried out differently.

From a technical point of view, while the alignment pipeline employed in PanoInserts worked reliably during our experiments, this solution suffers from a major limitation. The matching scheme adopted here matches a video (i.e. a 2D plane) to a face of a cube-map (i.e. a 2D plane). Even if practical, the proxy geometry used to approximate the panorama is far from optimal, and therefore this solution works well only when the video fits an entire face of the cube. When this is not the case, for example when the video is in transitions between faces, the alignment breaks, resulting in severe artefacts. Therefore, in the next chapter we will present a more general and reliable solution for the video to panorama alignment problem which uses a more precise proxy geometry (i.e. a sphere) as the target for the alignment. Similarly, to present a visually pleasant blending of the video with the static panorama, we used a crude colour

balancing scheme as outlined in Section 5.2.4. However, we are that more rigorous photometric-based colour correction techniques exist (see [YDMH99, Por03] for examples), and we reserve this aspect as areas of future improvements of the system.

Regarding the user evaluation, in Section 5.3 we described the process that lead us to design our experiment in the way it is outlined in this thesis, and we discussed the ecological validity of the tasks and critical parameters analysed in our study. However, it is important to note that different routes could have been explored during the investigation. For instance, while we believe that the tasks employed in our study are representative of common actions performed during multi-party remote meetings and remote assistance scenarios, we are aware that different aspects of remote collaboration could have been investigated. One interesting alternative would have been to investigate how often and how accurately users employed spatial deixis during their remote interaction. The problems with remote spatial understanding are usually manifested in the inability to point to, or reference objects in either local or remote environments. Indeed, real-world collaborative tasks are frequently performed through extensive usage of spatial deixis, which ground the interaction through referential statements and gestures made in relation to objects of common interest [Fil82]. Therefore, one interesting alternative to our evaluation would have been to record the participants dialogues while performing spatially-related tasks, such as remote manipulation of objects, and then, similarly to [FKS00, LHK+03], analyse how often spatial-references occurred. A different approach could have been to focus the study on small-scale objects manipulation (see [RBB06] for an example), and then analyse both the accuracy in the manipulation and the dialogues content. Finally, we could have also analysed post-experimental sketches of rooms in which the interaction took place, possibly asking users to either draw the layout of the room or to fill in a provided maps with the location of certain objects.

Given the alternatives outlined above, we believe that our experimental design could be improved with few modifications. First, we believe that adding dialogue analysis, especially if focussed on spatial deixis, could give a more in-depth understanding of the interactions that occurred during our tests. Similarly, we could have collected repeated measurements for each participant and task, minimising the novelty effect of our system. Hence, we suggest these modifications as possible extensions for future research.

During our study we compared PanoInserts to two broadly different VMC systems, which presented, theoretically, different degrees of spatial information. The reason for doing do so lies in the fact that we were interested in investigating aspects which are intrinsics of the visual representation. Thus in our experiment we build an ideal teleconferencing continuum in which the highly portable, but scarcely immersive webcam based video-chats represent the lower end of the interval, while the highly immersive, but scarcely portable fully-panoramic systems represent its end point. However, a different route could have been taken, albeit investigating different aspects of our system. The literature presents few similar systems to PanoInserts, the most notably being CamBlend by Norris *et al.* [NSQ12]. In a variation of our study we could have compared PanoInserts to CamBlend, concluding on how the two systems varied. However, this would not have answered our fundamental research questions of whether video in

context can improve spatial understanding and remote collaboration. Similarly, as our system is able to support several smartphones running in parallel to populate a static panorama with dynamic inserts, we could have investigated the effect of varying the number of phones during our tasks. However, given our research questions, it is critical to assess the quality of our fundamental approach without being diverted into assessing how this may change as the number of dynamic inserts increases. Thus, we decided to restrict our design to a single overlay. However, both variations of our study present interesting research points, which we hope to investigate in future work.

Finally, an interesting point to consider is the external validity of our study. We already discussed how we focused our experimental design on tasks which are representative of real-world usage (see Section 5.3). Following from that discussion, we believe that our results generalise well to others settings and scenarios, in which relatively large FoV contexts are employed. When this is not the case, and either the interactions taken into account or the visual representation employed are largely different from our study, we caution the reader from drawing conclusions from our results. In our study we find out that adding spatial context to canonical VMC systems can in turn greatly improve spatial understanding and benefit remote interaction. However, essential precondition for this is a relatively large FoV context which can augment the spatial references available to the users beyond what is achievable with standard video. Therefore, while we believe that the tasks chosen for the study well represent the real-world usage of the system, and thus results drawn from them can also generalise to other tasks, including remote space explorations, virtual tourism or even search and rescue scenarios, we are aware that the results obtained in our investigation are limited to videos+panoramic-contexts systems.

Another concern with the generalizability of the findings of our study is that the tasks focused on one particular goal (i.e., manipulate objects), ignoring the fact that during real-world meetings or interactions a variety of external factors can influence both the conversation and the performance. In the study participants were required to arrange objects in space, with minimal interactions with the confederates and no other distractions. This is unlikely in real-world scenarios, where dual interaction is often key for success. However, as we found strong evidence that videos in panoramic contexts can enhance remote collaboration, we believe that our study can contribute to corroborate previous work on similar subjects.

5.6 Chapter Summary

This chapter presented a user study to evaluate the effect of videos in panoramic contexts for remote collaboration. To conduct the study, we developed PanoInserts, a system allowing users to rapidly assemble a set of cameras to generate a panorama with live inserts for use in teleconferencing applications. After motivating the experimental aims in Section 5.1, the chapter described the system architecture (Section 5.2). The description then focused on the user study (Section 5.3.1), with a discussion on design, data collection, procedure, hypothesis and results. The chapter then ended with a discussion on the results (Section 5.5), analysing task performance, system usability, properties of the prosed spatial representation and implications of the experiment's outcome on the overarching theme of this thesis.

Results indicate that our system performs comparably with fully-panoramic video, and better than

webcam video conferencing in tasks that require a surrounding representation of the remote space. This suggests that our approach lies between fully-panoramic and webcam-based video both in terms of its technical characteristics and device accessibility, and also in terms of the richness of the conveyed spatial information that users can demonstrably understand and act upon. We demonstrated how a network of dynamically relocatable cameras allows users to capture dynamics and spatial relationships which would be hard to perceive otherwise. This is an important finding that shows how videos in panoramic contexts can help users building spatial maps of remote places and support spatiality, all fundamental properties of ICVE system, and thus of BEAMING.

In the results analysis we have also discussed issues relating to the problematic visual perception of panoramas due to varying distortion according to depth. To complement this discussion and further analyse this problem, in the following chapters we will introduce two additional experiments designed to investigate both hardware and software methods for displaying videos in panoramic context in a visually-intuitive manner that also promotes spatial reasoning.

Chapter 6

Experiment: Videos in Context for Spatio-Temporal Browsing

The biggest difference between time and space is that you can't reuse time.

Merrick Furst

In the previous chapter we have investigated the suitability of videos in panoramic context for teleconferencing. The experimental results showed that, when performing tasks that require a high level of spatial thinking, users can benefit from static panoramas augmented with a live inset recorded and streamed from within a remote location. Additionally, the experiment showed that users can intuitively understand and act upon panoramic representations of remote locations, even if these present the space in an unusual way (i.e. equirectangular projection). Nevertheless, the aforementioned user study investigated a single video inset. While this configuration allowed for a fair comparison with existing video-chat systems, it did not permit an investigation of the full potential of the proposed representation. As we will show later in this chapter, when multiple videos are combined together, more information on the dynamics and on the spatial properties of a remote environment can be inferred and consequently a richer visual representation of the remote location can be obtained.The videos in panoramic contexts representation then, is able to offer a unified view of an ensemble of videos that, when further grouped using a common spatial context, can greatly benefit a variety of activities, such as video-surveillance, virtual tourism, multi-parties video-conferencing or virtual exploration of remote environments.

Therefore, this chapter describes a second user study that investigates the effect of multiple videos in context on user spatial and temporal understanding of a remote scene. To perform the study we extended the focus+context paradigm presented in the previous chapter to create a video-collections+context interface that embeds several videos into a static panorama. To broaden the area of the study, and to conduct a wider investigation on the representation, we do not limit the videos to be streamed in real time, but rather we include in the collections videos recorded at different time. This, then, shifts the focus of the study from space only to space and time.

We call the developed interface *Vidicontexts* (see Figure 6.1 for an overview). To support the system, we built a spatio-temporal index and tools for fast exploration of the space and time of a video-collection,

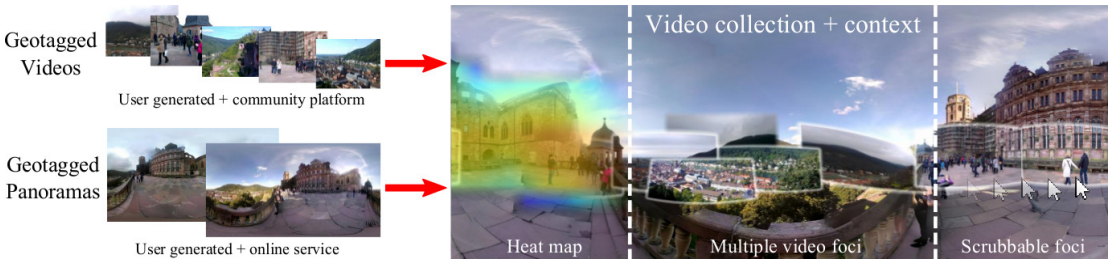


Figure 6.1: *Panoramas are widely available online, and more and more video content of these places is shared online. With these data, our video-collection+context interface visualizes the dynamic changes within a collection. The right-hand side shows our spatio-temporal index as a heat map (*left*), inlayed video foci (*center*), and fast search with spatial mouse scrubbing (*right*).*

and we investigated its usage and suitability for temporal and spatial related tasks. We compared the proposed interface with existing video browsing tools, analysing users performance, strategies and impressions. While the experimental design, user study and consequent data analysis presented in this chapter have been entirely carried out by the candidate, the development of the *Vidicontexts* system is the result of a shared development effort between the candidate and another developer – Dr. James Tompkin. Specifically, Dr. Tompkin was responsible for the development of the rendering and GUI modules of the system, while the candidate was responsible for the video alignment module. The remaining modules were developed in collaboration.

The remainder of this chapter is structured as follows. The motivations of the study, and consequently the system, are introduced in the next section. The chapter continues with technical implementation details of our system, including, video alignment, spatio-temporal index construction and allowed interactions. An evaluation of the system based on performance and timings is then introduced, followed by an user study addressing the fundamental implications for spatial and temporal perception over three video-browsing tools: a standard video browsing application (Apple's iMovie [App14]), the same application augmented with a panoramic view, and our system. We will show that our representation offers a highly performing solution in terms of spatio-temporal thinking, and that our system allows for a variety of interactions, which are not available on standard video-browsing tools, greatly enhancing users' performance. Finally, we will discuss implications of the users study results, exploring how the varying spatio-temporal forms of video-browsing are perceived and how they influence users when performing spatio-temporal reasoning tasks. A video showing the system in action, as well as additional informational material, can be found on the system's webpage[1]. Please note that some of the images reproduced here are extracted from the author's own work [TPS+13].

6.1 Motivation

The abundance and pervasiveness of mobile devices featuring built-in cameras has enabled people to document several aspects of their lives and change the way they communicate. This resulted in a fast diffusion of mobile videoconferencing, and in an ever increasing number of videos of places around the world. With geotagging, it is very easy to assemble a video-collection containing many videos of

[1] http://gvv.mpi-inf.mpg.de/projects/Vidicontexts/

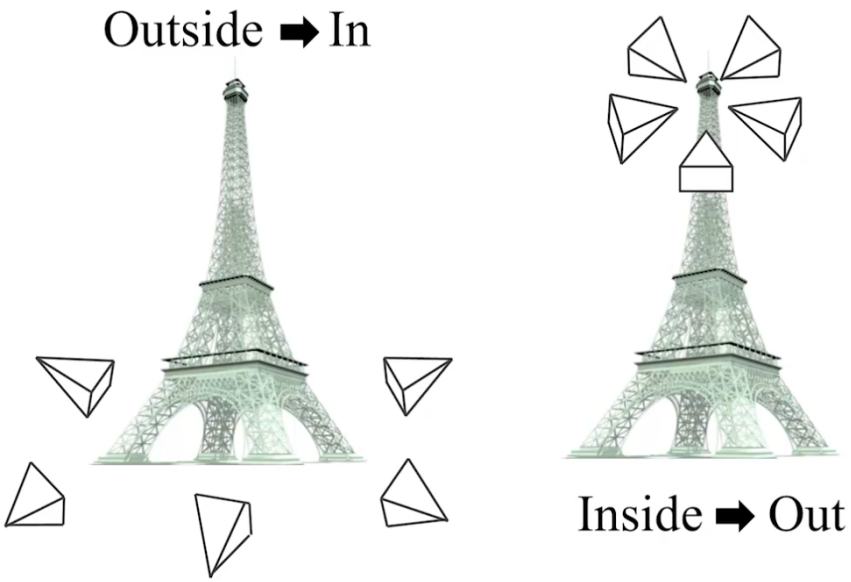


Figure 6.2: *Video-collection types.* Left: *In outside→in video-collections videos are captured from different viewpoint, but all look at the same place of interest.* Right: *In inside→out video-collections videos are all captured from roughly the same location.*

the same location spanning a period of time. Such a collection can capture both the moment-to-moment dynamics of a location, the comings and goings, and its temporal evolution across days, months, seasons, or years. As such, video-collections of places show contrasts and changes in our world. However, exploring these dynamic changes within places is difficult for users as existing interfaces do not explicitly connect the spatio-temporal content and display it within a unifying context. For example, a virtual tourist wishing to explore the dynamic events taking place over time in a famous square can only see videos in isolation, and has no easy tools to search within the space or time of the place. The exploration of the collection, then, results in a quite tedious process which may involve watching the same video several times in order for the user to build spatial and temporal mind maps.

The current primary way of exploring a video collection is by searching through metadata such as name, description, rating, date, or popularity. This searching technique, while perfectly functional for finding music videos and clips from named shows, is much less practical when wanting to find videos of a place or an event where the search term is typically less descriptive. In this case, metadata searches do not exploit content similarities or useful additional data from sensors. This difficulty in providing content-based similarity interfaces is reflected in the work disseminated by the multimedia retrieval and indexing communities (such as ACM Multimedia and ACM International Conference on Multimedia Retrieval). In such venues, while few works explicitly target video browsing interfaces that exploit video similarities for retrieval and presentation [GSW11, SB11], the main focus remains on the algorithmic efforts of retrieval.

Hence, current mapping applications such as Google Maps [Goo07], while linking videos geographically and providing ways to find videos taken from the same place, do not explicitly relate the changes over space and time into a single view for easy comparison, and users must watch videos in

turn. Clearly, this interaction paradigm is not optimal, and it is far from the way we interact with the real world. In fact, our experience of implacement (i.e., the way we understand space), as Edward Casey has termed it, is one of understanding our situational location [Cas93]. This, as argued by Farman [Far14], is typically done in a number of ways. One important way of doing this is by orienting our bodies in a proprioceptive way, i.e. we understand space by relating our body's position in relationship to the people and objects around us. In other words, we understand space and we orient ourselves in it by simultaneously establishing relationships between us and the objects and people around us, and by "centring" the space around our body.

State-of-the-art research systems for video-collection browsing, such as Unstructured Video-based Rendering [BBPP10] and Videoscapes [TKKT12], try to find visual links within videos that all observe the same content either at the same time or across different times. However, often the contents of a geotagged video-collection captured from the same place will not visually match because the videos all look out from approximately the same spot: we define these contents as "inside→out" (see Figure 6.2, right). For instance, two videos of a touristic vista might take in side-by-side views but never intersect. Further, for many interesting places it is impossible to "go around" and we can only "look around", such as atop the Eiffel Tower in Paris of from within Trafalgar Square in London. This forbids the application of existing vision-based matching systems which rely on cameras in different positions which converge to a common scene: we define these contents as "outside→in" because the cameras surround the subject (see Figure 6.2, left). As such, currently it is difficult to structure, relate, and explore inside→out video-collections, which however better mimic the way we understand and position ourselves in space.

To solve this problem, we introduce Vidicontexts, a system that embeds videos into the common context of a panoramic frame of reference. Vidicontexts extends the focus+context paradigm presented with PanoInserts, and enables the simultaneous visualization of individual videos as multiple foci, and through the context allows the exploration of how videos are spatially and temporally related even though there might be no direct visual match between them. Starting from the experimental findings obtained with PanoInserts, we develop Vidicontexts with the firm belief that its visual representation can alleviate the difficulty of spatially and temporally exploring inside→out video-collections. However, contrary to PanoInserts which tackles the problem of teleconferencing, the system here presented handles offline video browsing. The reasons for this are twofold. Firstly, we are interested to explore the effect of videos in panoramic context on tasks that require a high level of spatial and temporal reasoning. Second, we are interested in developing a system that can handle a large video-collection (i.e. more than 20 videos), at interactive rates. While the latter is theoretically possible also with streamed video, in practice receiving and encoding dozens of high definition videos simultaneously would add a substantial computational overhead to the system, jeopardising the interactivity of the communication. However, studying the effect of the proposed representation on temporal thinking can only be achieved with pre-recorded videos. For these reasons we restrict our system to work offline ans therefore, even though Vidicontexts borrows some fundamental concepts from PanoInserts, the two system are quite different.

In Vidicontexts we align geotagged video from mobile devices to a panoramic context using a

combination of orientation sensor data (if available) and time stamps and feature-based registration. Omnidirectional panoramas exist for many places from online street mapping platforms, and recent work enables accurate pairing of geolocated images and panoramas [KWO10]. Further, as already described in Section 2.3.1, panorama stitching is a common easy-to-use application for a variety of devices (including mobile devices). These sources provide readily available contexts for our video-collections. In general, any task that requires spatial or temporal reasoning would benefit from our system. A user might browse a collection of videos to locate object in space/time, follow videos, infer temporal changes, highlight captured regions, filter and isolate video instances that belong to a particular time span or spatial bounds; broadly, relate videos within a collection. Sport, museum, cultural sites, social events, surveillance, and tourist videos could be browsed and analysed. If extended to support live video streaming, our system could serve crucial tasks such as remote assistance, rescue or medical inspections.

Vidicontexts can be classified as a focus+context system [BGBS02], as it presents a novel way to link and explore collection of videos in their original context. Ideally our representation is able to convey more spatial and temporal information than conventional video-browsing tools. This is due to the abundance of videos linked together, and the information that users can retrieve from this. In the previous chapter we established that a single video in panoramic context can convey spatial information that users can understand and act upon. Nevertheless, proving that this property directly extends to the representation here presented is not trivial. Specifically, the abundance of visual stimuli and the unconventional panoramic representation of the context could potentially confuse the users. Therefore, to test this, we decided to run a user study. Linking back to the initial hypothesis presented in Chapter 1, the aim of the study was to understand whether contextualising large video-collection through a spatio-temporal index and with the aid of static panoramas can help user improve spatial and temporal understanding of remote places (i.e. H1 and H4). To do so, we designed two tasks which involved spatial and temporal reasoning, and we studied how users' performance varied across different browsing modes that included our novel representation, a standard video-browsing tool, and the same tool augmented with a panorama.

6.2 Architecture Overview

Vidicontexts takes as input a panoramic image and a collection of videos with time stamps, GPS data, and orientation sensor data. We first track and align all videos within the panorama, which yields a sequence of homographies for each video. Next, we build a spatio-temporal video index for exploration. Finally, we provide an interface to explore the collection of videos within their panoramic context.

The Vidicontexts interface (Fig. 6.3) presents the context in either look-around perspective projection or as a full equirectangular map projection with an infinitely-rotating canvas. The interface is divided in four main operational areas: video selection in the top, video foci playback in the middle, video playback controls in the lower left corner and context manipulation in the lower right corner. The user is free to pan, zoom, and smoothly switch between perspectives. Videos can be visually followed, or the context can be locked to follow individual videos. We also provide standard video playback controls.

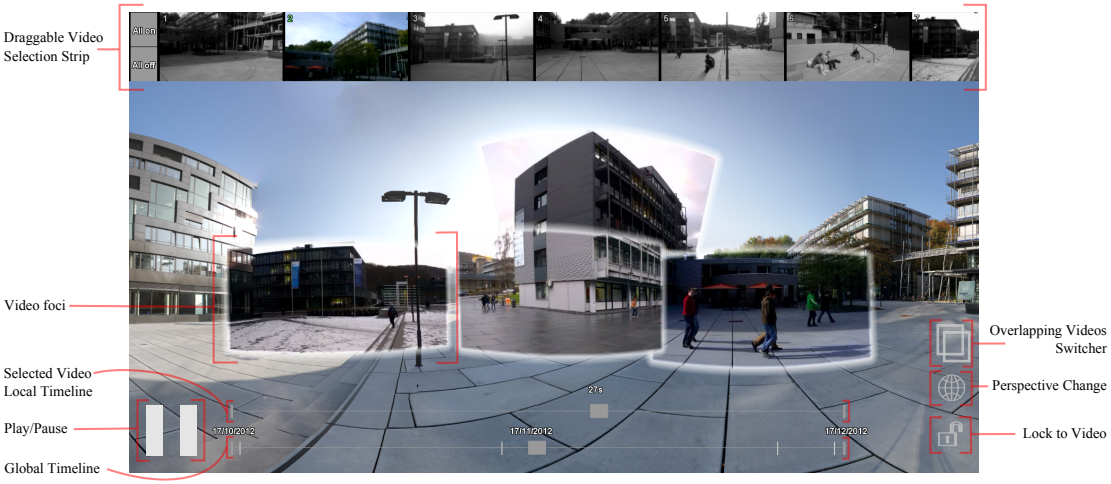


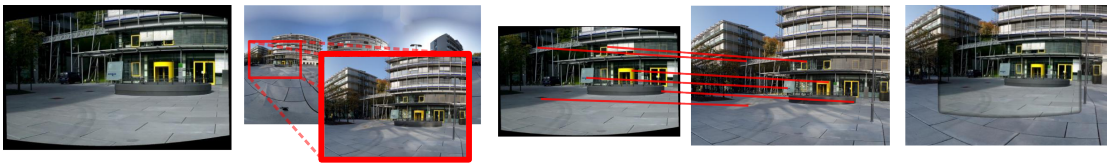
Figure 6.3: *The Vidicontexts interface. Different months of the fall season – rainy October, cloudy skies in November, snow in December – and dynamic objects are added to the summer scene context.*

6.2.1 Capture and Context

The panoramic contexts used in Vidicontexts are the same employed in PanoInserts. As such, they can be acquired from online repositories such as Google Street View [Goo07], panoramic cameras, and digital single-lens reflex camera (DSLR) stitches, or from user-assisted tools included in many mobile devices. Although any suitable source could be used, to create the material for the results showed in this chapter and for the user study we used Microsoft Photosynth [Mic08] on smartphones and Microsoft Research's ICE [Mic12c] for stitching photos from a DSLR. Next, we captured several example video-collections ourselves from roughly the same location as the panoramas, returning to the same locations over time. We captured a variety of environments, including college grounds, a castle vista, a modern courtyard, a neo-classical quad and an indoor hallway. Figure 6.11, which is placed at the end of the chapter, shows the captured environments. We used Samsung Galaxy SII [Sam11] and HTC OneX [HTC12] smartphones to capture both video, GPS location, and orientation data. To record the data we developed an application for the smartphones to obtain readings from the camera, integrated accelerometer, gyroscope, and magnetometer sensors via the Android API [Goo08a]. This camera orientation estimate provides an initial registration to the panoramic context. For online panoramas, pairing geotagged videos to geotagged panoramas can be difficult when GPS data is inaccurate. For the results showed in this thesis, we manually picked the pairing panoramas. However, if automatic pairing is desired, we assume existing work to pick the closest geographical panorama from an online repository [KWO10].

6.2.2 Video Alignment

Orientation data provides only approximate video alignment to the context. Accurate spatial localization is made difficult by *a*) hand-held video capture with jitter; *b*) time changes between context and videos, causing lighting changes, static and dynamic object changes, and broad scene appearance changes from seasonal variation; and *c*) the computational cost of alignment traded-off against the need to handle collections of videos. While developing PanoInserts, we decided to leverage an alignment solution that used a cube as a target geometry. However, as already discussed in Section 5.2.2, such solution has



(a) Spherical projection (b) Localization with sensor data (c) Feature-based alignment (d) Panorama composition

Figure 6.4: *(a) Videos a projected into a spherical coordinate system and (b) localized using orientation sensor data. Within this localization, SIFT features are extracted and robustly matched to estimate an alignment (c) for compositing (d).*

severe limitation, especially when the video moves across faces of the cube, and therefore we decided to improve a different alignment strategy for our platform. Despite variations in capture pose, we assume that the spherical panorama is a good proxy geometry for the scene, and we align the video frames to the spherical panorama using sensor- and feature-based image alignment (Fig. 6.4). Even though a direct alignment of planar surfaces to both quadric [CZ98] and non-regular surfaces [CKV+09] is possible, such process is computationally expensive and error prone. Therefore we decide to transform the perspective video frames beforehand the alignment to facilitate it. As a consequence, contrary to PanoInserts, in which marker-based tracking is used only to bootstrap the systems or as a fall-back solution, in Vidicontexts when gross tracking is available from the phone's sensor, it is systematically used to speed-up the alignment process and to make it more reliable.

Spherical Projection. We transform perspective videos into spherical projection with focal length meta-data and pitch and roll orientation data from our smartphones, following the principles expressed in [BK01]. If the focal length, pitch, and roll estimates are accurate, and if there is no parallax, then the spherically transformed video frames would be related to the equirectangular panorama by a translational model; however, due to errors in these estimates, we allow more freedom in the alignment transformation by using a homography model.

Feature Extraction. During our system development, we tested a variety of image descriptors and feature detectors, including DAISY [TLF10] dense descriptors and FAST [RD06] and SIFT [Low04] features. Contrary to [KWO10], in which DAISY image descriptors are employed to localise images in panoramas, we found SIFT features to have the best performance, and therefore we employed them. Initially, we extract and store the features from the panorama and subsequently, we extract them from each spherically-warped video frame. As feature extraction is a frame-independent task, we parallelise it.

Sensor-data based Localization. We localize video frames approximately within the panorama using orientation data. Given this, we only match panorama features to video features within a bounding box 20% larger than the approximate localization. This makes the alignment process more reliable, as it significantly reduces matching time and false matches. For videos with no meta-data or sensor readings, we perform an initial robust feature-based match between the panorama and the video to discover approximately the focal length, pitch, and roll angles.

Homography Estimation. With four or more matches between frames and panorama, we can estimate a homography between each video frame and the panorama using the gold standard algorithm [HZ04],

which employs RANSAC refinement [FB81] over an initial set of matches. For further refinement, we use the estimated homography to find inliers from the initial set of matches and re-estimate the homography using inliers only, as suggested in [Far05]. This refinement step is repeated for three iterations in our implementation, but different heuristic could also be used. As we have a strong expectation for a translation transformation, we perform conservative homography outlier rejection and remove homographies that are not projective or that have a large skew factor. For completeness, a MATLAB function for homography validation is reported in Appendix E, Figure E.2.

Estimation of Missing Homographies. With outlier rejection, it is possible that no good homography is found for short sequences of frames. We approximate these missing homographies by interpolating between two valid homographies. With a neighbouring valid homography as a starting point, we accumulate sensor orientation changes until we find a valid homography end point, and then integrate the resulting error over the length of the missing sequence.

The estimation works as follows. We project the corner positions of frame rectangles within panorama space using sensor rotation data. Let us denote projected corners of frame i as TL_i, TR_i, BL_i, and BR_i (see Figure 6.5). Using these corner positions we estimate the angle between neighbouring frames after orientation-based projection. We also compute the x and y translation of these projected frames as the difference between the centroids of these corners after projection.

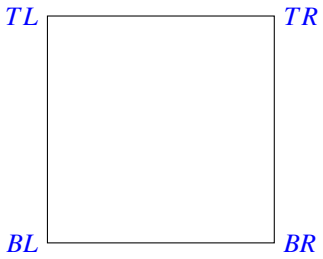


Figure 6.5: *Projected corners of a frame used for our homographies estimation.*

Let us denote the neighboring frames as i and $i+1$, the translation as tx_i and ty_i and the angle between edge vectors e_i and e_{i+1} as θ_i. We can estimate θ_i using the dot product:

$$e_i = TR_i - TL_i \tag{6.1}$$

$$\theta_i = \arccos\left(\frac{e_i \cdot e_{i+1}}{\sqrt{\|e_i\| \, \|e_{i+1}\|}}\right) \tag{6.2}$$

The dot product in Equation 6.2 only gives us a positive θ_i. To discover the sign of the angle, we need to define a normal to the surface N, and use the cross product:

$$\theta_i = \begin{cases} -\theta_i, & \text{if } N \cdot (e_i \times e_{i+1}) < 0. \\ \theta_i, & \text{otherwise.} \end{cases} \tag{6.3}$$

In our 2D case, $N = [0,0,1]$, with e vectors zero padded in z. More simply, this operation reduces to checking the sign of $(e_{x,i}e_{y,i+1}) - (e_{y,i}e_{x,i+1})$.

We use the signed θ_i, tx, and ty to compute the corresponding affine transform between every pair of neighbouring frames:

$$H_{\theta_i} = \begin{bmatrix} cos(\theta_i) & -sin(\theta_i) & tx_i \\ sin(\theta_i) & cos(\theta_i) & ty_i \\ 0 & 0 & 1 \end{bmatrix} \tag{6.4}$$

The homography between frame $k+j$ and the panorama can be then estimated as the cumulative multiplication of the latest known homography matrix H_k and the estimated affine projection matrices H_{θ_i}, where $i \in [k+1, k+j]$.

$$H_{k+j} = H_k \cdot H_{\theta_{k+1}} \cdot ... \cdot H_{\theta_{k+j}} \tag{6.5}$$

Temporal Filtering. Since frame homographies are estimated independently, some temporal jitter remains due to small but independent alignment errors. Such jitter affects the visual quality of the registration, resulting in a distracting factor for the users. Hence, we bilaterally filter the frame corner positions over 30 frames in time to reduce temporal jitter. We modulate the contribution of each filter window position (temporal weight) by the image-space Euclidean distance from the center window position (range weight).

Let us denote the four corner points for frame i as P_i^k and the filtered corner locations as Q_i^k, where $k \in [1,4]$. Let us denote the centroid of the four corner points for frame i as P_i^c. The filtered locations are estimated using a bilateral filter:

$$Q_j^k = \frac{\sum_{i=j-T}^{j+T} W_t(i) \cdot W_s(P_i^c, P_j^c) \cdot P_i^k}{\sum_{i=-T}^{T} W_t(i) \cdot W_s(P_i^c, P_j^c)} \tag{6.6}$$

$$W_t(i) = e^{\frac{-i^2}{\sigma_t^2}} \tag{6.7}$$

$$W_s(P_i, P_j) = e^{\frac{-\left\|P_i - P_j\right\|^2}{\sigma_s^2}} \tag{6.8}$$

Here, the temporal filter W_t is the domain filter which ensures temporal smoothing. The spatial filter W_s is the range filter, which ensures that when the frame position difference is large (i.e. due to a sudden change in camera position or erroneous homography), it is not propagated through smoothing. We use the temporally filtered corner points Q_i^k to estimate inter-frame homographies, and we multiply these with the original homographies to produce a temporally smooth series of transformations.

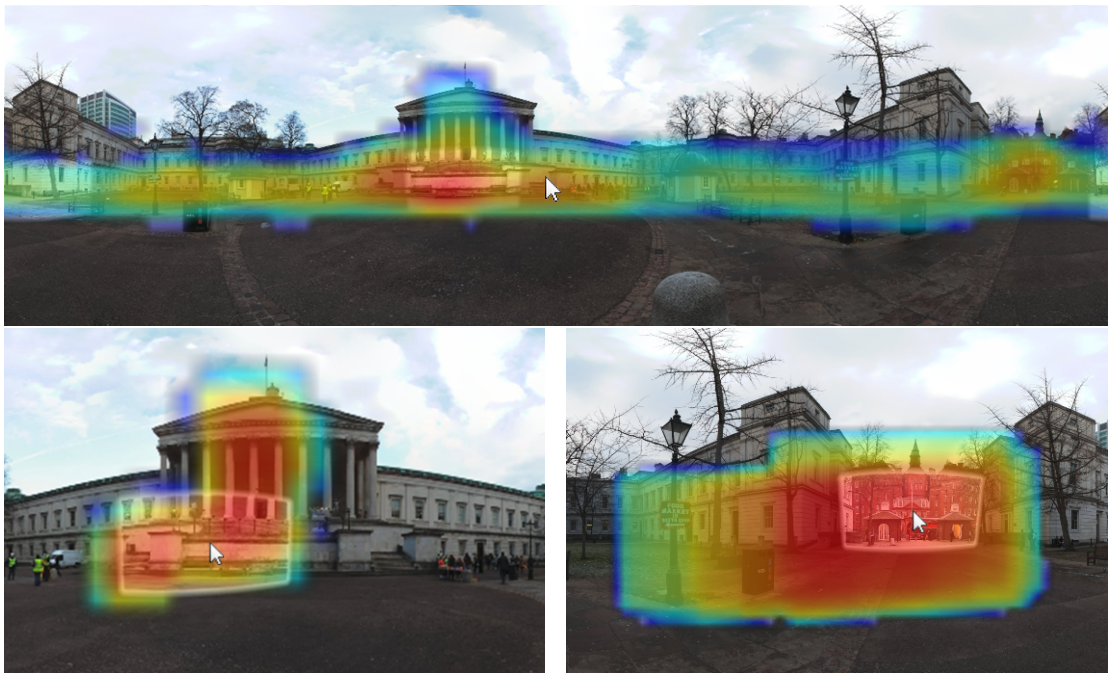


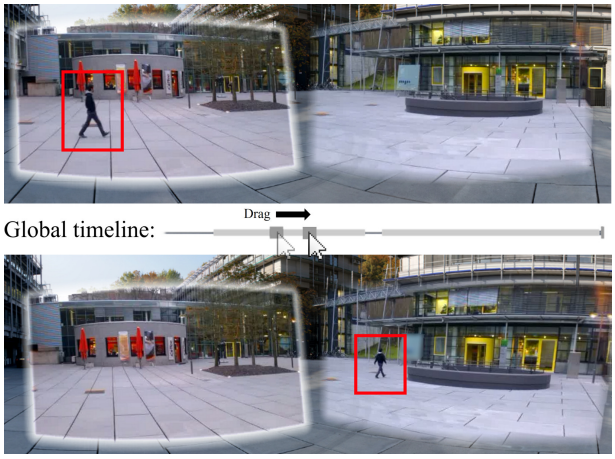
Figure 6.6: *The spatio-temporal index displayed as a heat map to show attention over the context. Red indicates areas where the collection focuses the most, blue where it focuses the least. This index allows quick spatio-temporal search and filtering of the video-collection. This is computed globally for the whole collection (*top*) and locally for each video (*bottom*).*

6.2.3 Spatio-temporal Index

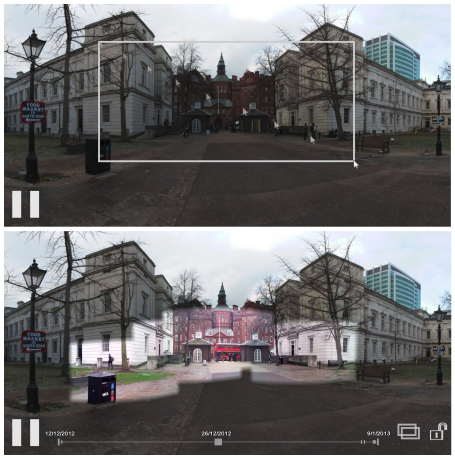
With video alignment, we can construct a spatio-temporal index. The index holds information on where and when each video intersects the context, and is the fundamental tool to allow a vast range of spatio-temporal interaction with the collection. From a conceptual point of view the index corresponds to the spatio-temporal mind-maps that an user browsing a collection of videos needs to establish to relate videos together. As such, the index acts as a valuable tool for the user to leverage his/her spatio-temporal thinking effort while browsing the collection.

We iterate through each video and intersect its per-frame bounds against a grid of cells which cover the panorama. The choice of grid resolution depends on the size of the dataset and memory constraints, with larger size resulting in higher computational times. In our implementation we set the cell resolution of this index to be $\sim 100 \times 50$ cells, which gives a moderate 40 pixel spatial precision across the panoramic context. Each grid cell stores the spans of frames per video which intersect it. The index is computed only once per dataset, and it is then stored in a binary file for later usage.

The spatio-temporal index can be visualized in many ways depending on the application. We choose to render the index with a gradient such as a heat map (Fig. 6.6). With this, users can see which regions of the context held the most "attention" among the videos, and our spatio-temporal interaction tools then allow these videos to be found quickly. Selecting individual videos shows a per-video index which defines the spatial extent of the recording. Heat maps for specific index queries can be generated, such as video attention for a historic time span. Other visualizations would be possible, such as altering the saturation of the panoramic context locally for when it is important not to overlay further graphics onto



(a) *The global timeline allows control over all videos at once and enforces temporal ordering. Here, two contemporaneously captured videos which never intersect are displayed synchronously and in context, allowing easy identification of the movements of a man (red square).*



(b) *The spatio-temporal index makes searching simple. The user draws a bounding box over the region of interest, and our interface lays out all intersecting videos. The dynamic objects, such as the red bus, bring the scene to life.*

Figure 6.7: *Temporally-driven and spatially-driven interactions.*

the scene, or by displaying the path of the video by rendering a line joining the center-most grid cells along the path. Arrows on the line can show the progression of time, and color with a gradient can show where the video lingers.

6.2.4 Interface and Interaction

The Vidicontexts interface (Fig. 6.3) presents standards controls to interact with both the videos in the collection and the context. Additionally, our video-collection+context interface provides tools for spatio-temporal interactions which make interacting with it more interesting.

Temporally-driven Interactions. Each video has its own local timeline which appears when the video is selected. Unlike a normal video player, adjusting the timeline affects both the dynamic content within the video and the spatial position of the video in the context, and this provides a quick way to check the spatial extent of a video. This enables new applications: by adjusting the timelines of different videos and by setting A→B loop markers, the user can compose a novel arrangement of the videos within the context to highlight spatio-temporal changes. By setting the markers the user can infinitely loop through videos, perhaps focusing on a certain area of the scene or a given time-frame.

As we have timestamps for each video, we also show a global timeline which displays the temporal extent of the video-collection. Adjusting the ends of the timeline filters the video-collection, for instance, to isolate videos from a particular day or month. As such, the global timeline effectively acts as a temporal filter on whole collection. The global timeline slider synchronously adjusts the playback of all videos in the collection, and allows the visualization of events which share the same time but otherwise have no visual overlap (Fig. 6.7(a)). Such relations are difficult to explore when the videos are seen out of context, as it would require to repeatedly inspect the collection to isolate events happening at the same time in different videos. If multiple panoramas captured from the same position are available and timestamped, then the global slider also switches between them. This shows temporal changes in the

Dataset	# Videos	# Frames	Alignment	Index
College Grounds	15	30,426	6hr 10min	40sec
Castle Vista	9	17,460	3hr 33min	25sec
Modern Courtyard	11	21,518	4hr 26min	30sec
Neo-classical Quad	20	26,635	6hr 16min	55sec
Indoor Hallway	6	4,152	36min	16sec

Table 6.1: *Computation times for alignment and spatio-temporal indexing (100×50 cells) for our datasets. Examples of the datasets are reported at the end of this chapter in Figure 6.11.*

context: for instance, in the seasons or in the built environment.

Spatially-driven Interactions. Temporal scrubbing has a spatial equivalent: By dragging the mouse over the panoramic context, the user can spatially drag individual videos or all videos at once, providing a fast way to localize many videos at once. As videos are not guaranteed to visit all locations in space, they scrub to their nearest position. The extents of each video individually and of all videos combined can be shown by visualizing the spatio-temporal index (Fig. 6.6), and this helps guide spatial exploration.

We also provide area-based spatio-temporal filtering (Fig. 6.7(b)). By dragging a box over the context to describe a region of interest, the user queries the spatio-temporal index for sequences of frames which intersect the region. This is a very fast way to "collage" an area of the context with video. Practically, the area-based filtering allows the user to isolate regions within the panorama to narrow the focus on the collection. Such tool can be extremely useful in situations when the user needs to monitor a certain region of the environment, e.g., in a surveillance task.

6.2.5 Performance timings

Vidicontexts provides an interface for offline browsing of video-collection, and as such some pre-computation is required to build the spatio-temporal index. We process videos independently and, as feature extraction is frame independent, our technique is embarrassingly parallel. Video alignment was computed on an Intel Xeon 8 core 2.40GHz PC; see Table 6.1 for computation times. All panoramas are 4000×2000 pixels, and all video frames are 1920×1080 pixels. Our alignment code is written in MATLAB and C++ (mainly using OpenCV [Wil99]), though GPU-accelerated matching algorithms may speed this up. The software runs on PCs running Windows XP or higher, Ubuntu 11.10 or higher and uses Java OpenGL [Jog10] for rendering. The computation time for the spatio-temporal indices is also included in this table, and this performance scales linearly with the total number of cells.

The computational performance of our interface is defined by the number of videos visible. The rendering cost is minimal as we need only apply a homography to a pre-warped video and its feathered matte; however, the video decompression cost is large. Our implementation supports approximately 3 1080p HD videos at frame-rate at once. To cope with more videos, we store a reduced resolution version at a quarter scale, and only switch to full resolution if the user zooms in. While modern CPUs, such as the Intel Quick Sync on Sandy Bridge or later CPUs, contain hardware to decompress 5 or more videos at once [Shi11], it is difficult to use this as our video format must support fast and exact seeking.

6.3 User Study

Vidicontexts facilitates spatio-temporal exploration and comparison within video-collections. While this is straightforward to understand and demonstrate, measuring whether our proposed representation provides significant benefits to perform spatio-temporal related tasks is non-trivial. In Section 5.3 we have introduced a discussion on why it is important to evaluate novel technologies through ecologically valid experiments, and how choosing the right critical parameters to focus on during the evaluation is not straightforward. The evaluation of Vidicontexts is no exception to this, and therefore in the rest of this section we will motivate the decision made while designing our experimental study.

In the user study conducted on PanoInserts we have established that a single video in panoramic context can improve over standard webcam video for tasks in which spatial reasoning is required. The representation proposed in Vidicontexts adds new information compared to the PanoInserts case as it presents a) a visual stimuli that comprises of a large number of videos and b) a representation that combines space and time elements in context. This results in a richer visual stimuli than the one offered by PanoInserts, albeit a larger amount of information to process ad understand.

The user study presented in this chapter can be then considered as an incremental improvement over the experiment presented in the previous chapter. As such, the implications of the study results help answering some of the research questions introduced at the beginning of this thesis. Specifically, the user study will address whether multiple videos in panoramic context can be used to convey spatial and temporal information of a remote place, and if this is beneficial to users' spatial and temporal thinking; Additionally, the experiment will help us assessing if such representation can be easily understood and acted upon and if, given its design, it will also help us in understanding whether our system provides significant benefits over existing video-collection browsing interfaces.

As already argued in Section 5.3, selecting truly representative critical parameters, and identifying the type of activities that a system is designed to support, are crucial steps to design an ecologically valid study. In our case, we designed Vidicontexts with the goal of facilitating applications where users need to obtain an in-depth understanding of both spatial and temporal relationships between several videos or cameras. Such applications span several domains, and in here we give few examples. Surveillance is an important application which commonly produces data from cameras mounted to pan and tilt heads, and this exactly fits our scenario of video+context. Critical tasks might include reviewing videos over time and space for suspicious behaviour, or reviewing videos over time and space to identify and localize a person of interest. Another interesting application domain for our representation is virtual tourism, whose industry we image will implement new systems to display videos of the time and space of a place, requiring then exploration interfaces for these applications. For instance, our experimental setup well-models a system where users upload their own personal videos of a famous place, to be explored as part of an online collection of all videos uploaded of that place within a context say, an enhanced, user-driven Google Street View. This might even be extended to include treasure hunts or puzzles games, similar to existing panoramic games such as GeoGuessr[2] or Myst 3[3]. Additionally, our representation naturally

[2] https://geoguessr.com - last accessed 09/12/2014
[3] http://presto.yune.me/presto/titles.html - last accessed 09/12/2014

fits panoramic telepresence applications, as the ones introduced in Section 5.3.

Therefore, motivated by these example applications, our study analyses the videos in panoramic context representation when used to perform two tasks that require participants to infer spatial and temporal information from a video-collection. The two tasks focus on counting people located in a particular area, and track people that crosses a certain region of interests, respectively. We compare Vidicontexts with *iMovie* (Figure 6.8), which offers a chronological browsing window and a resizeable timeline for fast preview, and against iMovie augmented with a panoramic image available for reference (*iMovie+pano* henceforth).

Before running the experiment as it is documented in this chapter, the experimental design was refined over several iterations. Initially, we listed a number of possible tasks to evaluate the effect of video-collection+context on spatio-temporal-related tasks. It soon emerged that tasks chosen to measure performance should represent general actions performed regularly by users interacting with video-collections. We identified common actions while exploring a place, including looking for objects/people in space and in time, following dynamic events within the place, and identifying when changes happen within specific times or areas of the place. Therefore, we selected two tasks that involve counting and tracking events, in our case the comings and goings of people, within several videos. These tasks offer two reliable metrics which a) mimic common tasks performed when browsing a video-collection, and b) can be extended to multiple system interfaces for comparison. In addition, we excluded possible tasks which would be trivial with one interface over another (e.g., in our interface, to find all videos which intersect part of the panorama). The resulting tasks are exemplars for real interactions which allow us to assess different systems and validate spatial and temporal understanding.

Hypotheses. In both tasks, we measured the completion time and accuracy expressed as errors in the people counts, and we collected the results of usability and tasks related questionnaires. Given our premises, we expect accuracy to vary with the sophistication of the spatio-temporal representation, and so we expect Vidicontexts to be more accurate. In turn, we expect iMovie+pano to be more accurate than iMovie alone. For completion time, we expect performance to vary according to the spatio-temporal controls available, and so we expect our system to require the least time. We expect all three conditions to score above average (75%) on the usability questionnaire. Finally, we expect Vidicontexts to obtain the highest score for the task-related questionnaire as we believe this is directly related to task performance.

6.3.1 Method

Participants

30 unpaid participants from the staff and student population at our university performed both tasks using one system each for a between-subjects design for the system independent condition, and a within-subjects design for the task. Participants were recruited via e-mails and other forms of messaging. While we did not filter the study population for handedness and eyesight, we ensure gender balance was respected. Additionally, the participants were randomly assigned one of the three systems, within which the order of the two tasks was alternated to minimize the influence of learning effects. All subjects were introduced to their assigned system and to the tasks, and there was no mention of the overarching goal

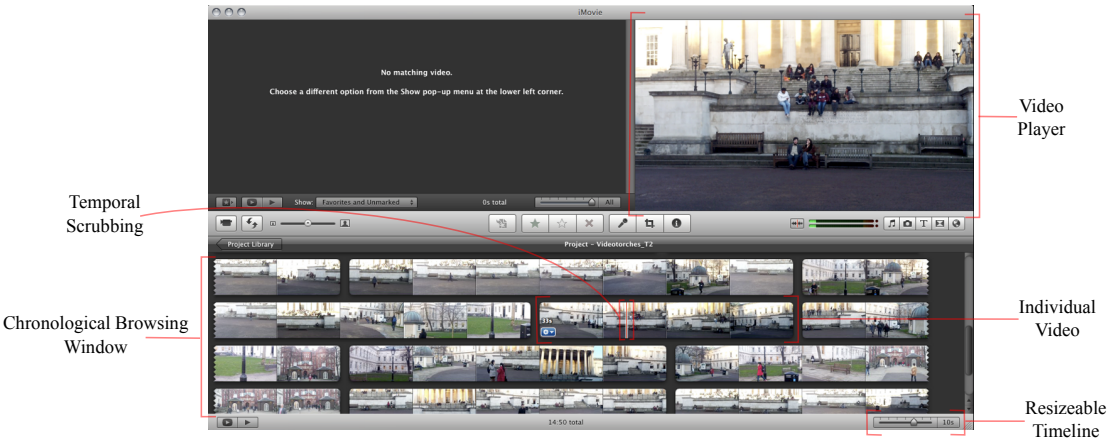


Figure 6.8: *The iMovie interface used in our user study.*

of the study. All participants were familiar with editing in general, and all received training with their system.

Design

We wish to assess the accuracy with which participants can correctly obtain a spatial and temporal understanding of a collection of videos. Hence, in both tasks, we measured the completion time and accuracy expressed as errors in the people counts. To do so, after each task we asked participants to fill a form with their answers. Following the experiment, participants completed the standard System Usability Scale (SUS) questionnaire, which gathered subjective assessments of usability of the three systems, for the full set of questions, please refer to Brooke [Bro96] or Appendix E. Additionally, participants were asked to answer height task-related questions (see Appendix E and Table 6.3 for a list of questions), and their impression on the system was also recorded.

Procedure

As discussed before, our study was split over two tasks that require participants to infer spatial and temporal information from a video-collection. For each of the three conditions investigated we used the same exocentric non-immersive display (flat desktop display - Dell U2410) with mouse control over a cursor (Belkin Optical Ergo mouse).

The first task, which we will call the *people counting* task henceforth, requires participants to browse twenty videos from the *neo-classical quad* dataset (Figures 6.6 and 6.7(b) and fourth row of Figure 6.11) and identify the number of different people who sit on a set of benches. Videos differ in length and cover the entire horizontal area of the environment. As different videos could depict the same person, or show a person sitting near the areas of interest, a participant could potentially make twenty erroneous counts. The maximum number of errors was manually counted.

The other task, which we will call the *people tracking* task henceforth, asks participants to review six videos from the *modern courtyard* dataset (Figure 6.7(a) and third row of Figure 6.11) and track the number of different people who cross between two buildings. Here, the videos never fully track a person and do not overlap, so multiple synchronous videos must be analysed to obtain the correct result. Videos

Condition	People Counting		People Tracking	
	Error	Time	Error	Time
iMovie vs. +pano	0.958	0.916	0.968	0.898
iMovie vs. Ours	0.040	0.017	0.049	0.014
+pano vs. Ours	0.107	0.023	0.012	0.005

Table 6.2: *Significance (p-values) for each task and condition combination for both error and time to complete. Green values are statistically significant ($\alpha = 0.05$).*

differ in length, but they all cover a similar area of the environment (approximately 125° horizontally). A participant could potentially make twelve erroneous counts (again manually counted).

Each participant performed two different tasks using the same system, with no time limit. Participants could use all features of each system, e.g., in iMovie and iMovie+pano the built-in video scrubbing and thumbnail expansion. Figure 6.8 shows the iMovie interface used as a comparison condition. For the readers not familiar with the system, a detailed description of the interface is reported in Appendix E. Before starting the experiment, each participant was given a detailed description of the system's interface and features, and as much time as they liked to familiarize before the task. The participant then was presented with a form in which a brief for each task was reported. We ensured participants understood the tasks, and, if necessary, answered their questions. Each task was conducted in series, with no interaction between the participant and the experimenter. Following both tasks, the participant completed two questionnaires and his/her impressions on the system and study was recorded.

6.4 User Study Results

The primary dependent measures of interest used for both tasks were the accuracy expressed as errors in the people counts and the time taken to complete the task. Initially, for statistical analysis a 3×2 (system × task) mixed Analysis of Variance (ANOVA) was computed using SPSS [IBM09] to analyse each of the dependent variables. System was a between-subject factor, while task was a within-subject factor. A significance level of .05 ($\alpha = 0.05$) was used for judging the significance of effects. No samples were removed from our analysis, as all the participants successfully completed their tasks.

6.4.1 Accuracy

The counting error results shall be considered first. Figure 6.9 presents a summary of the accuracy results. It shows mean number of errors committed and standard deviation, as well as the significance, for each task and condition combination.

For the *people counting* task, participants had fewer errors for Vidicontexts ($M = 1.4$, $SD = 1.17$) and iMovie ($M = 4.9$, $SD = 3.87$) than for iMovie+pano ($M = 5.5$, $SD = 5.58$). For *people tracking*, participants had fewer errors for Vidicontexts ($M = 0.9$, $SD = 1.19$) and iMovie ($M = 6.2$, $SD = 5.94$) than for iMovie+pano ($M = 6.8$, $SD = 1.17$). Results of statistical analysis found a main effect of system on accuracy ($F_{(2,27)} = 12.836$, $p < 0.001$). However, no main effect of task on accuracy was reveal ($F_{(1,27)} = 0.294$, $p = 0.592$). The interaction between system and task was also not significant ($F_{(2,27)} = 0.216$, $p = 0.807$).

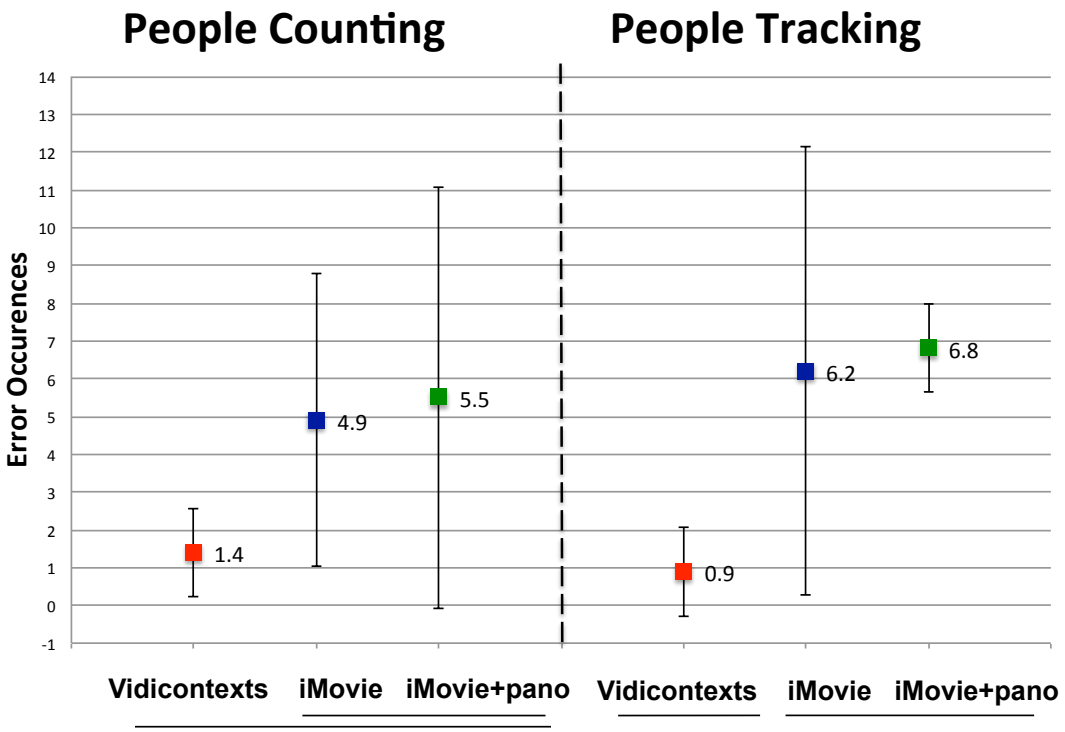


Figure 6.9: *Mean error occurrences and standard deviation for the three systems in both tasks. Conditions jointly underlined are statistically similar.*

To further unpack the effect of the between-subject factor (i.e., system) at each level of task, we computed two ANOVA (one per task) with the system used as the single factor and counting error as the dependent variable, with post-hoc Games-Howell tests for pairwise significance tests (cf. Table 6.2).

For the *people counting* task, a non-significant main effect of the system used was found, even if a statistical trend can be observed ($F_{(2,27)} = 3.09$, $p = 0.06$). Post-hoc analysis revealed non-significant differences between Vidicontexts and iMovie+pano ($p = 0.107$), albeit large differences between mean and standard deviation values leading to a statistical trend, and significant differences between our system and iMovie ($p = 0.04$). A main effect was not found between iMovie and iMovie+pano ($p = 0.958$).

For *people tracking*, the system used showed a significant main effect ($F_{(2,27)} = 5.08$, $p = 0.013$). Post-hoc analysis revealed significant differences between Vidicontexts and iMovie ($p = 0.049$), as well as between Vidicontexts and iMovie+pano ($p = 0.012$). No main effect was found between iMovie and iMovie+pano ($p = 0.968$).

6.4.2 Completion Time

We will now analyse the dependent variable completion time. Figure 6.10 offers an overview of the completion time results. It shows mean and standard deviation, as well as the significance, for each task and condition combination.

For the *people counting* task users took less time for Vidicontexts ($M = 469$ sec., $SD = 121.85$ sec.) and iMovie ($M = 662$ sec., $SD = 144.29$ sec.) than for iMovie+pano ($M = 688$ sec., $SD = 195.76$ sec.). For *people tracking*, Vidicontexts obtained the lowest mean time ($M = 373$ sec., $SD = 94.13$ sec.),

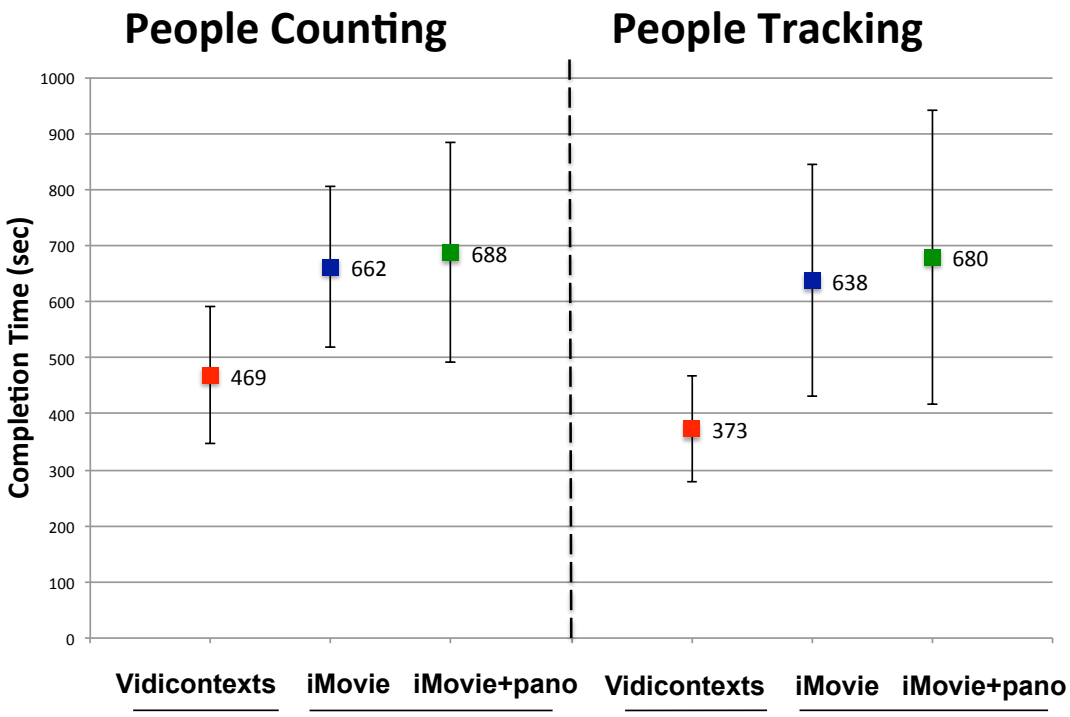


Figure 6.10: *Mean time to complete and standard deviation for the three systems in both tasks. Conditions jointly underlined are statistically similar.*

followed by iMovie ($M = 638$ sec., $SD = 207.13$ sec.) and iMovie+pano ($M = 680$ sec., $SD = 262.81$ sec.). Results of statistical analysis found a main effect of system on completion time ($F_{(2,27)} = 7.796$, $p = 0.002$). However, no main effect of task on completion time was reveal ($F_{(1,27)} = 2.169$, $p = 0.152$).

Also in this case, the interaction between system and task was not significant ($F_{(2,27)} = 0.885$, $p = 0.424$). Hence, to further unpack the effect of system at each level of task, we again computed two ANOVA (one per task) with the system used as the single factor and time to complete as the dependent variable, with post-hoc Games-Howell tests for pairwise significance tests (cf. Table 6.2).

For the *people counting* task, the system used was a significant factor ($F_{(2,27)} = 5.60$, $p = 0.009$). Post-hoc analysis revealed a significant difference between Vidicontexts and both iMovie ($p = 0.017$) and iMovie+pano ($p = 0.023$), and non-significant differences between iMovie and iMovie+pano ($p = 0.916$). For *people tracking*, the system used was again a significant factor ($F_{(2,27)} = 7.16$, $p = 0.003$). Also for this level of task, post-hoc analysis revealed a significant difference between Vidicontexts and both iMovie ($p = 0.014$) and iMovie+pano ($p = 0.005$), but non-significant differences between iMovie and iMovie+pano ($p = 0.898$).

6.4.3 Questionnaires

For the usability questionnaire (SUS), only our system scored above average ($SUS = 77.5$), followed by the iMovie+pano ($SUS = 62.75$) and iMovie ($SUS = 59.5$) conditions. Following the SUS classification technique of Lewis et al. [LS09] (letter-grade ranks varying from A to F), Vidicontexts is a Rank B system, while both iMovie and iMovie+pano mode are Rank C systems. Rank A systems have many promoters, who will definitely use and recommend the product. Rank B systems have a fair number of

Task-related question	iMovie	+pano	Ours
Easy to complete tasks	2.3	2.6	4
Understood video orientation in space	3.5	3.9	4.7
Understood relative video position	3	3.8	4.4
Understood space-time video overlap	2.8	3.8	4.3
Understood temporal order of videos	1.5	2.1	3.4
Environment representation confused	3.2	3.5	1.7
System has enough functions for tasks	3	2.5	4.4
#videos made remembering things hard	3.9	4.2	2.6
Overall mean	2.375	2.62	3.86

Table 6.3: *System mean scores for the task-related questionnaire. The response scale varies between 1 and 5. The scale for negative questions was reversed for mean computation.*

promoters, who are likely to use and promote the product. All other ranks will only have detractors.

For the task-related questionnaire, both iMovie and iMovie+pano conditions performed poorly, with mean scores of $M = 2.375$ and $M = 2.62$ respectively. Our system scored higher on this questionnaire, with a mean of $M = 3.86$. Further analysis revealed that there is a significant main effect of system used ($p = 0.001$), and post-hoc analysis reveals significant differences between Vidicontexts and iMovie ($p = 0.001$), as well as between our system and iMovie+pano ($p = 0.003$). Table 6.3 presents the mean score for each system and question.

6.4.4 Participants Comments

Following the experiments, we recorded participants impressions. Regarding our system, no negative remarks were registered. One participant commented that *"keyboard shortcuts to fine-grain navigate through timelines"* where required as *"sometimes one needs more control"*.

Impressions on both iMovie conditions were more negative, with a predominant feeling that the two systems were unnecessary cumbersome to complete the tasks. In particular, one iMovie's users commented that *"I saw some guy walking, but when I wanted to compare, I couldn't find him anymore"*. Another participant reported that *"the tasks would have been much easier if the videos were ordered temporally"* and that *"if you resize the thumbnails it's easy to forget which video you were last looking at. Annoying if there are so many videos that you need to make smaller to fit in the workspace"*. Interestingly, one of the iMovie+pano users commented that *"the task would be easier if the videos would be temporally and spatially aligned. Ideal would be a system that would save thumbnails of people the user clicks on, so that he can remember by himself whether a people he just sees entering or leaving the area of interest entered or left the area in another sequence before."* Some of the suggestions in this comment resemble some of Vidicontexts functionalities.

6.5 Discussion

6.5.1 Tasks Strategy and Performance

The results from our user study reveal insights into the way participants were able to spatially and temporally perceive and act on information presented in the varying systems. In both tasks, Vidicontexts

provides greater accuracy while obtaining the lowest time to complete, and this agrees with our initial hypothesis. The significant reductions in error and time to complete the tasks confirms that our spatio-temporal representation combines necessary information to reduce task complexity over iMovie. Analysing tasks strategy helps explaining this last concept.

While analysing the counting task we note that both iMovie conditions' users need to spatially locate the videos before counting people, as only particular regions are of interest. In contrast to this, Vidicontexts' user need only to count people as the videos are already spatially located. This reduction in complexity allows the user to perform only the task's essential action. This suggests that the proposed representation can encode spatial and temporal information that people can intuitively understand and act upon. Analysing the user task strategy confirms this further: in the counting task, for iMovie and iMovie+pano, users first expanded the video thumbnails timeline to spatially locate each video in turn, and then either used normal playback tools or scrubbed through the videos as thumbnails to count people. For Vidicontexts, participants could exploit the spatial alignment and only needed to search in time. Here, most used the local video timeline tool. This "natural" way of exploring the environment reduced the time to complete the task, and, in general, required less information to be autonomously inferred and processed by the user.

A similar trend can be found while analysing the people tracking task. Here, the task requires temporal and spatial alignment, and this increased cognitive load presented a challenge to users. In both iMovie conditions, users had to replay parts of the collection several times before answering. One user in both iMovie conditions struggled to accomplish the task at all, and generally participants from these two conditions struggled more than participants from our condition. For Vidicontexts, the global timeline maintains temporal alignment, and so this was frequently used by the participants in this task. No users struggled to complete the tasks with our system: the context representation combines necessary information and reduces task complexity. As a results participants' effort was greatly reduced, as users could simply look at temporally and spatially aligned videos, and easily track people crossing different areas of the environment. This, in turn, reduced the time to complete and greatly improved accuracy.

6.5.2 Usability

For both iMovie conditions, user task strategy was to first expand the video thumbnails timeline to obtain an idea of where each video pointed, and then either to use the playback tools or to scrub through the videos as thumbnails. For the people tracking task, users played parts of the collection several times before answering. In general, participants from the iMovie and iMovie+pano conditions struggled more than Vidicontexts users, with one user in both iMovie conditions finding difficult to accomplish the task at all. One user in both the iMovie and iMovie+pano conditions struggled to accomplish the task at all, and generally participants from these two conditions struggled more than participants from our condition. The limited functionalities of the two systems have been pointed out by several users in their post-study interviews. Obviously, this has had an impact on both performance and usability assessment, which emerges from both the usability and the task-related questionnaires.

For Vidicontexts, most participants used the local video timelines to accelerate video localization

in the panorama. The global timeline was frequently used by the participants in the tracking task, but rarely used for the counting task. No users struggled to complete the tasks with Vidicontexts. These positive aspects of the systems have been often praised by the participants. Additionally, users quick familiarised with the novel Vidicontexts interface, as video information is presented in a similar way to real life environments. This can help explain why users response to our system was also positive in terms of usability and desirability, as suggested by the much higher questionnaire scores for our interface than for both iMovie conditions.

Analysing individual questions further reveals that participants considered our system the best tool to convey spatial and temporal information within the video-collection, that they perceived our representation as less confusing, and that they thought our tools were more useful for exploration tasks. Additionally, participants agreed that, for tens of videos, our system improved recall. Such positive reception of our system demonstrates that our interface has a high level of usability and desirability, and suggests that the proposed representation, beside being functional, is also easily understandable.

Finally, we observe a general trend in the preferred panorama projection. To complete the tasks, 80% of the population assigned to our system used equirectangular projection. This finding, in accordance with recent work by Mulloni et al. [MSD+12], shows that participants thought the 360°-at-once projection conferred more spatio-temporal information than the geometrically-correct perspective projection. We assume that users did not want to be constrained to a limited FoV for localization tasks.

6.5.3 Conclusion and Limitations

The experimental work described in this chapter reveals fundamental insights on the quality and benefit of video-collections in panoramic context for spatially and temporally browsing. In particular, results showed that when a collection of videos is presented in a way that reproduces the original spatial and temporal arrangement of the videos, browsing such collection and inferring information from it requires significantly less cognitive load that would otherwise be required with standard video browsing tools.

Linking back to the initial hypothesis presented in Chapter 1, the aim of the study was to understand whether contextualising large video-collection through a spatio-temporal index and with the aid of static panoramas can help user improve spatial and temporal understanding of remote places (i.e. H1 and H4). The importance of the results presented here, then, is manifold. First, we can establish that the proposed system presents a powerful interface which can greatly improve over existing (and largely diffused) video browsing tools. Besides positive performance, our system has had a favourable reception, demonstrating its suitability for spatio-temporal related tasks. Additionally, we proved that videos in panoramic context can convey spatial and temporal information of a remote location that standard video representations cannot replicate. This is an extremely important finding which reveals insights on the videos+focus paradigm. First, the variety of visual stimuli offered by the representation do not confuse the users, but rather help them spatially and temporally organise the collection. Second, environment's landmarks, which are a fundamental part of the context, are often exploited by the users to infer spatial relationships between videos of the collections, but also between people and objects within the videos.

Combining these findings together then, helps us addressing one of the main research question of

the thesis. In Chapter 1, we hypothesised that contextualising large video-collection through a spatio-temporal index and with the aid of static panoramas could help user improve spatial and temporal understanding of remote places (i.e. H1 and H4). We can confidently say that the results presented here confirm these hypothesis. Videos in panoramic context help users understanding spatial and temporal relationships within remote location and video-collections, with a visual representation that people can intuitively understand and act upon.

Hence, if we further extend the results presented here with the one introduced in Chapter 5, it is clear that the videos in context paradigm offers a beneficial representation for both video-conferencing systems and offline video browsing. Therefore, we can conclude that, as providing panoramic context is a special case of the general problem of aligning content to world model, the proposed representation offers a valid crutch to provide space-time exploration of remote environments, confirming our initial hypotheses (i.e. H2).

While the system and study presented in this chapter helped us answering some of the research questions of this thesis, we note that our implementation of both the system and the study represents only one of the viable routes that we could have followed. However, in particular for the user study, we believe that we focussed our design on aspects which closely reflect real-world usage and which are thus representative of real problems to investigate. In its current form the system can only handle videos which are captured in an area that overlaps the center of the panoramic context. While loosening this constraint is possible, in practice implementing it is non trivial, as it would require a different proxy geometry for the alignment (i.e., an arbitrary shape, three-dimensional model). Nevertheless, the problem of extending our visual representation to full 3D models is appealing and challenging, and opens up an interesting direction for future development (please refer to Section 9.3 for further discussion on this point).

From a technical point of view, Vidicontexts is limited to offline browsing. While this is partially due to the nature of the activities that the system was designed to support (e.g., surveillance, virtual tourism, remote exploration), it is also true that current solutions employed for aligning the videos to the static panorama are infeasible for real-time usage. While this is not an issue for video replay, it is problematic for interactive applications, in which our system could be used, for instance, to convey live events in real-time (e.g., a music festival). However, registering dozens of videos in real-time is certainly a non trivial tasks, which would require a combination of engineering effort and algorithmic improvement. Hence, we reserve to improve this aspects in future works, as outlined in Section 9.3.

Regarding the user evaluation, we have already motivated earlier in this chapter (cf. Section 6.3) the choices made while defining its design. However, different routes could have been taken at different stages of the implementation. For instance, different tasks could have been selected, and different parameters analysed. One choice of task could have been to recreate a virtual tour of a remote environment, and then asses both the spatial and temporal understanding of users. We could have asked the users to sketch a map of the place they had just explored, situating objects on that same map and answering questions related to events that happened in time. Similarly, we could have focused our evaluation on

a surveillance-like task; participants would have had to browse through a variety of videos of a specific event, and later either identify a specific event, or reply to a set of questions focused on both objects locations and specific events. Also the overall link between videos could have been exploited. In a different choice of task we could have shown a variety of spatially and temporally related videos to the users, and then ask to sort them manually with and without the aid of our system. However, it is important to note that the majority of these tasks would have been much easier when completed through Vidicontexts than through the other systems. As this is partially related to the nature of the systems involved and to the novelty of our representation, we decided to opt for more general tasks which would not obvious favour our system design.

Besides tasks, we could have also designed the experiment to take into account our system novelty, and trying to minimise this factor while comparing the performance of the same users across different system. To this end, we could have collected repeated measures for each system, and designed a within-subject study in which the same participant would have conducted a set of tasks using all the different systems analysed. However, we decided to opt in our study for a between-subject design for the system condition, completely avoiding any learning effect, and simply focussing on system's response. Nevertheless, a future study could investigate this different design.

Finally, linking back to the outcome of our study, we argue that our experimental design allows us to generalise the results to a variety of real-world usage and system configurations, including in situ augmented reality exploration, immersive surveillance setups and virtual entrainment systems. However, and similarly to what concluded for PanoInserts (cf. Section 5.5.4), we caution the reader from overgeneralising our results. As our study only took into account offline video browsing, remote meetings applications or VMC system cannot be directly linked to our results. However, we note that the tasks selected for our study represent typical actions that users perform while browsing large video collections, especially when looking for specific events, people or objects in space. As such, we believe that extending Vidicontexts to handle live streams and support remote collaboration may present an interesting future research direction, in which the benefits of both PanoInserts and Vidicontexts could be merged together to convey highly spatially-enhanced VMC.

6.6 Chapter Summary

This chapter presented a user study that investigates the effect of videos in context on user spatial and temporal understanding of a remote scene. To conduct the study, we extended the focus+context paradigm introduced in Chapter 5 to create a video-collections+context interface. The interface, which we call Vidicontexts, embeds several videos into a static panorama, enabling spatio-temporal browsing. To complement the investigation presented in previous chapter we did not limit the videos to be streamed in real time, but rather we included in the collections videos recorded at different time, shifting the focus of the study from space only to space and time combined.

The chapter has been structured as follows. After motivating the experimental aims in Section 6.1, the chapter introduced the system architecture (Section 6.2). The chapter then focused on the user study (Section 6.3), with a description of design, data collection, procedure, hypothesis and results. We then

introduced a conclusive discussion on the results (Section 6.5), analysing task performance and strategy, system usability, properties of the prosed spatial representation and implications of the experiment's outcome on the overarching theme of this thesis.

Results indicate that our system performs better than typical video browsing tools in tasks that require to infer spatial and temporal properties of a remote space, providing significant benefits to accuracy and time taken in such localization tasks. Additionally, our interface is preferred both in general by the SUS and specifically for our tasks. This suggests that our approach offers a richer visual representation in terms of space and time than standard videos. This is an important finding that shows how providing panoramic context makes spatio-temporal tasks easier and faster. Supported by these results and previous experiments, we can conclude that videos in panoramic context can help users building spatial and temporal maps of remote places to enhance spatiality, a fundamental properties of ICVEs, and thus of BEAMING.

In the last two chapters we investigated visual properties of videos in panoramic context. We evaluated two different interfaces, and inferred properties of the visual representation that can be extended to the broader class of video+focus systems. However, the video-collection+context representation fits display and interaction devices beyond desktop environments, such as tablets, spherical displays, and HMDs. These devices map the panorama to both virtual and real spatially-located spheres. Different displays, then, provide different real and virtual geometries, and this might impact how users relate to and perform with the panoramic context. The next chapter will investigate this aspect with a final user study.

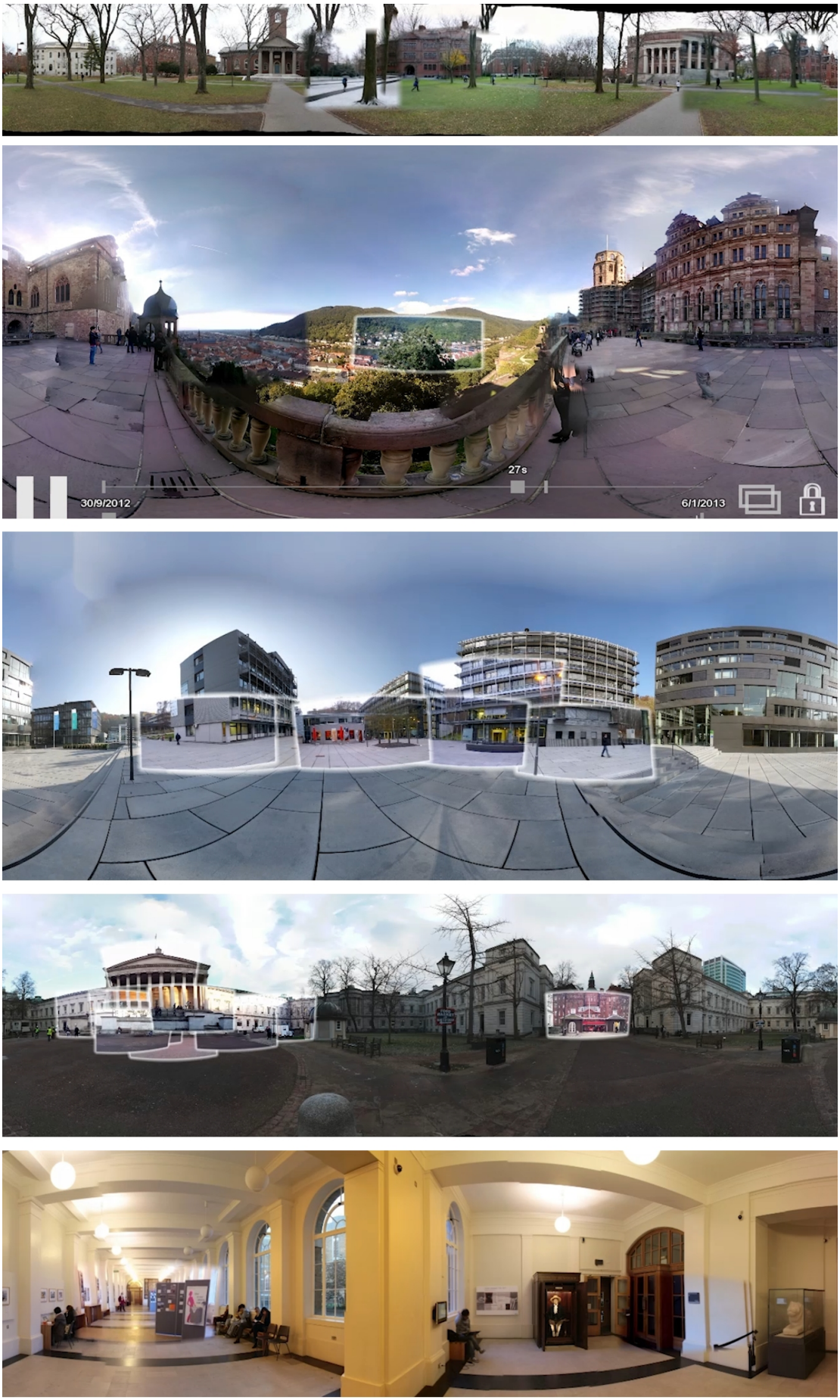


Figure 6.11: *Datasets captured for the results showed in this thesis and for the users study. From top-down: College Grounds (sensors only alignment), Castle Vista, Modern Courtyard, Neo-classical Quad and Indoor Hallway.*

Chapter 7

Experiment: Immersive Display Effect on Videos in Panoramic Context Tasks

A display connected to a digital computer gives us a chance to gain familiarity with concepts not realizable in the physical world. It is a looking glass into a mathematical wonderland.

Ivan E. Sutherland

In previous chapters we investigated visual properties of videos in panoramic context, and we established the positive aspects of such representation with respect to spatial and temporal thinking. We note though that the video+focus representation also fits display and interaction devices beyond desktop environments, such as tablets, spherical displays, and head-mounted displays. These devices map the panorama to both virtual and real spatially-located spheres, creating an immersive representation of the context. Different displays, then, provide different real and virtual geometries, and this might impact how users relate to and perform with the panoramic context. In this chapter, we present a user study whose aim is to analyse the effect of display type on user's spatial and temporal perception. Conceptually, the results presented here can be considered as complementary to the ones introduced in previous chapters. Indeed, the study focuses on the effect of different displays on the video+focus representation, rather than on the representation itself.

Similar investigations have been performed by virtual reality researchers. Immersive displays have been shown through virtual reality studies to potentially increase performance in 3D spatial reasoning tasks. Nevertheless, such studies have considered only 3D virtual environments. We note though that panoramic images and video lie between 3D environments and 2D images, as the user can be surrounded by the world but receives no parallax cues. Therefore, motivated by previous studies, we want to establish whether the potentially positive effect of display immersion in 3D environments can be extend to panoramas. To this aim, we adapt the Vidicontexts interface (see Chapter 6) and conduct a user study to discover whether display type affects spatio-temporal reasoning across desktop monitor, tablet, and HMDs.

The remainder of this chapter is structured as follows. The motivation of the study is introduced in the next section. The chapter then introduces the design space and implementation details of the interface

extension to novel displays. The user study is then introduced, and results are reported. Finally, we discuss how participants responded to the different interfaces, how they approached the reasoning tasks on each of the displays, and how they evaluated the displays in usability and task-related questionnaires. We combine this information and formulate implications for designing panoramic imagery systems. Please note that some of the images reproduced here are extracted from the author's own work [PTP+14].

7.1 Motivation

Virtual reality research has established that immersive displays, such as HMDs, can improve user performance in tasks that require a high level of spatial reasoning or in tasks that mimic the real world [PCS+00, MJSS02, TGSP03]. Conclusions from these studies suggest that less immersive exocentric displays (e.g., screen based, viewing the world from outside) are less performant than egocentric displays (e.g., HMD-based, viewing the world from inside) when users must reason about 3D virtual environments, and this has implications for applications such as games, telepresence, and scientific visualization.

Panoramic images and video are now common, with the world quickly being mapped at street level by companies and tourists alike. However, while the wide-spread use of existing panoramic imagery applications and the novelty of panoramic video applications is apparent, no research has yet touched on the effect of different display devices on these upcoming panoramic video imagery applications. This problem is brought into focus when we consider that some tasks which involve panoramic video imagery are performance critical, such as rescue services telepresence or security surveillance review. In previous chapters, we have demonstrated how panoramas can be beneficial to applications where spatial understanding of the scene is critical. We focussed our studies on panoramic imagery as this representation has hybrid visual properties that make it particularly interesting for our studies. On a spectrum between 3D virtual environments and 2D images, panoramas lie somewhere in between — a 360° panorama can surround a user, but the scene has only spherical geometry and is effectively flat. The user cannot move from their point of view and so does not receive parallax cues.

While this form of representation cannot be considered as three dimensional, some of its visual properties are closely related to the ones investigated in virtual environments. Therefore, to complete our investigation on panoramic imagery used as contexts, we have decided to study whether the immersive display's benefit for spatial reasoning in VE extends to hybrid 2.5D panoramas. In Chapter 1, we hypothesised that the level of immersion of a display type can be a significant factor on users spatio-temporal thinking, affecting the eventual beneficial properties offered by the video+context representation (i.e. H5). However, to our knowledge, the research community has not addressed this question yet. Its answer though has implications for many applications which require viewing, interacting with, and reasoning about panoramic imagery.

We explore this question with a user study investigating the impact of display type on reasoning about events within panoramic video scenes. To measure spatio-temporal reasoning performance, we use the Vidicontexts interface introduced in Chapter 6, and we employ it across three displays which sample interesting points within the immersive displays design space: 1) flat desktop displays, which

are exocentric; 2) mobile tablet displays with orientation tracking, which are free to rotate and act as windows into the world — these are egocentric but not immersive; and 3) HMDs with orientation tracking, which are both egocentric and immersive. This range of displays covers both common display types in desktops and tablets, and more unusual HMD displays which, with their recent affordability, are becoming more common.

Importantly, there is no clear application boundary which limits each display type, and each instance can potentially serve a variety of applications which require spatio-temporal reasoning: desktop displays could be used for surveillance applications and event monitoring, tablet devices could be employed for virtual tourism, while HMDs could be adopted for immersive visualization and telepresence. Therefore, when coupling the rich visual representation offered by the video+focus paradigm with each display type, it is not immediately clear which combination is the most beneficial. Immersive displays may improve spatial awareness, but their "immersion" may be confusing or overwhelming when combined with the numerous visual stimuli conveyed by our representation. Similarly, flat display may be more familiar to users, but at the same time it may limit the space awareness of the system, reducing tasks' performance.

To conduct the study, we adopted the same two experimental tasks introduced in previous chapter. We decided to replicate the two tasks as they mimic fundamental actions performed when exploring and reasoning about panoramic imagery, and so act as a proxy for other possible applications. These tasks require reasoning about the identity and whereabouts of people across space and time within panoramic scenes, and so mimic general localization, recognition, and tracking actions.

Our user study investigates the display effect on the previously unexplored hybrid 2D space of panoramic imagery and complements the results presented in the Chapters 5 and 6. Furthermore, different displays support different panoramic projections [MSD+12], and each display type forces a change of pointing interface [PSP93] — any display adaptation involves a trade-off, and our study has implications for this design space. Through our multi-display adaptation, and with the results of our user study, we discuss how participants responded to the different interfaces, how they approached the reasoning tasks on each of the displays, and how they evaluated the displays in usability and task-related questionnaires. We combine this information and formulate implications for designing panoramic imagery systems, and so hopefully aid future research and development in this field.

7.2 Vidicontexts Adaptation and Displays

The 360° nature of the imagery in Vidicontexts allows us to compare relevant display types. Furthermore, visualizing multiple video foci within the same panorama meets our need for engaging reasoning tasks. Video foci are captured on tripods or handheld video cameras, and so are free to rotate within the panorama and capture action across the space of the panorama. Video foci are also captured at different, potentially overlapping time spans, and so we can include temporal reasoning tasks too. With these foci, we can ask participants to follow or track objects, such as people, and so we can ask them to reason about the identity and whereabouts of people in space and time.

Vidicontexts presents video foci within a panoramic context (for a detailed description of the system

Properties	Desktop	Tablet	HMD	Spherical	CAVE
Input					
Mouse	✓	×	×	×	×
Touchscreen	✓	✓	×	✓	×
Joypad	✓	×	✓	✓	✓
Eye-track	!	!	!	!	!
Hand-track	!	!	!	!	!
Display					
Display size	24”	11”	7”	16” diam.	120”
Resolution	1080p	1080p	≈500x600 per eye	≈1024x768 per sphere	1080p per wall
Hor. angle	$\approx 50^\circ$	$\approx 25^\circ$	$\approx 100^\circ$	$\approx 90-135^\circ$	$\approx 180^\circ$
Immersive	×	×	✓	×	✓
Modes					
Egocentric	×	✓	✓	×	✓
Exocentric	✓	×	×	✓	×
Perspective	✓	✓	✓	!	✓
Equirect.	✓	!	!	!	×
Flipped space	×	×	×	!	×
In-situ	×	✓	×	×	×

Table 7.1: *The potential design space of display scenarios. Green marks the chosen display/input combinations, representing combinations likely to be found in practice. Exclamations mark interesting points discussed in the text.*

please see Section 6.2). It is important to note that this interface is able to adapt with minimal changes to different display types. First, the interface allows equirectangular map projection with an infinite-pan canvas, and look-around perspective projection. This allows the interface to adapt to exocentric displays with map projection, and to egocentric displays with orientation tracking for perspective projection. Both projection modes allow zooming into the scene, which should allow participants to overcome resolution differences between displays.

Displays. Choosing which displays to evaluate from the large number of possible configurations is tricky as each display type has different properties which might not be directly comparable. Trying to normalize these conditions is difficult. Instead, we choose a systems-level approach, where we try to compare systems which would most likely be used in practice. While this makes it harder to directly compare, instead, it allows us to see the impact of design decisions on participant behaviours with common systems. To illustrate the possible configurations to evaluate, we summarise them in Table 7.1. Within this design space, we choose to compare three display scenarios which provide both interesting points in the space and practical systems (Figure 7.1):

Desktop An exocentric non-immersive display, with mouse control over a cursor (Figure 7.1, *left*). The desktop runs the original Vidicontexts interface as presented in Chapter 6.
Hardware: *Dell U2410 with Belkin Optical Ergo mouse*.

Tablet An egocentric non-immersive display, with perspective orientation control through tablet rotation, and touchscreen control replacing a cursor (Figure 7.1, *center*). Similarly to [JKC12], our tablet interface performs perspective projection camera control through the device’s orientation

Figure 7.1: *Different display modes used for the study: (*left*) Desktop display with mouse; (*center*) Tablet with rotation and finger orientation controls; (*right*) HMD with head orientation and joypad cursor controls.*

sensors, allowing the user to physically rotate the device to navigate the context (Figure 7.2). In this way, the real proxy geometry of the scene is maintained as the user explores with a virtual window. A simple button press locks the orientation rotation and returns control of the virtual camera to touch.Additionally, front camera face tracking provides zoom control.

Hardware: *Acer Iconia W700*.

HMD An egocentric immersive display, with perspective orientation control through head rotation and joypad control over a cursor (Figure 7.1, *right*). The HMD is a binocular stereo device; however, we effectively make the display monocular by rendering views of a monocular panorama at infinity. Additionally, we render the graphical user interface (GUI) so that it follows head rotation at a fixed-depth into the world, with a cursor which moves only within the plane and bounding box of the interface.

Hardware: *Oculus Rift Dev. Kit with Xbox 360 wireless joypad*.

Bowman et al. [BDR+02] demonstrate that HMDs are a recommended choice when users require strong spatial orientation, outperforming CAVE-like systems. Coupled with their rarity in everyday life, we do not include CAVEs (large projection systems with head-tracking) and instead use a HMD for our egocentric immersive case. We also reject tablets physically located in the real world at the same location as the panorama, because there is no comparison for other display types. One interesting alternative is spherical displays, where a world captured from inside-out in a panorama is viewed outside-in looking onto the sphere. This has the effect of flipping spatial relations, where rotation around the sphere reveals imagery in the opposite direction to expectation. This is not necessarily a problem for spatio-temporal reasoning; however, while this would be interesting to test, we do not as it is a very rare display in practice. As both tablets physically located in the real world and spherical displays present interesting points of discussion, which however go beyond the scope of this chapter, we include a discussion on their usage in Appendix F.

Inputs. A change of display often brings with it a required change in input device, making the direct comparison harder. For example, an HMD with physical rotation is difficult to couple with a tethered pointing device, and a handheld tablet makes holding other devices difficult. As we take a systems-level approach, we choose points in the design space where all three display scenarios have different cursor controllers which are the most common input mechanisms for these devices (being mouse, touchscreen,

Figure 7.2: Top: *The tablet interface is free to rotate along all axes in space to provide a virtual window.* Bottom: *Front camera face tracking provides zoom control.*

and joypad). While a change in input device across experiment conditions can be a potential confound – and should certainly treated and acknowledged as one – we note that our investigation does not focus on input efficiency or any aspects that can be directly affected by that, but rather investigates the effect of display type on spatio-temporal reasoning. Therefore, we acknowledge the fact that a change in input device may emerge as a confound, but we argue that this aspect does not play a major role in our investigation, and hence should not affect our results.

Nevertheless, to our knowledge the literature has no strong conclusions about the absolute effectiveness of these pointing mechanisms, and performance is task and device dependent. There is some evidence to suggest that joypad input has reduced throughput to mouse input (0.69–0.33x bits/s) [SM04]. There are no wide surveys yet of touchscreen and mouse throughputs, but some touchscreen studies have suggested equivalent or faster movement times than with a mouse [SS91, FWSB07, JVS13], others that mouse input outperforms touchscreen input when the task requires a single point of contact [MCN94], but also that touchscreens potentially decreased accuracy [SMS09]. In principle, it would be possible to design eye- and hand-tracking systems which are suitable for all for these display types (see Table 7.1); however, these technologies are still nascent and uncommon, and a consistent integration across displays would be difficult.

As we change input device across displays, we state these important points: First, that interface

interaction time is insignificant compared to the expected task completion time. Second, that although our displays have varying angular extent, none have interface elements which fall below the critical angle of difficulty identified by Song et al. [Son12]. Third, that across our three displays, we ensure that the layout of GUI elements remains consistent by both making the GUI independent of the panoramic view — the GUI moves with your head — and scaled to the display size. To achieve this for the HMD, we render the interface elements onto a plane which follows head rotation at a fixed-depth into the world. For selection, a cursor moves only within this plane, and to mimic a screen, this cursor is bound to the interface elements in much the same way that a mouse is bound to the display's edges. Fourth, that we try to make world rotation amounts consistent with their displays to reduce the workload and error rate from the input devices. For both mouse and touchscreen inputs, a display edge-to-edge drag covers 360° degrees, and for the HMD there is a 1:1 mapping between head angle and panorama rotation.

Projections. For the exocentric desktop case, in systems and the literature, we have seen both equirectangular and perspective projection types commonly used, and no projection is considered optimal. As such, we decided to leave the choice to the user. However, equirectangular projections aren't consistent with egocentric view. For instance, even though the HMD is inherently egocentric, we could present an equirectangular panorama on a plane in a virtual desktop; however, this somewhat defeats our purpose for using an HMD. As such, we restrict tablet and HMD devices to use egocentric perspective projections.

7.3 User Study

In previous chapters we investigated, and established, the suitability of video+focus representations to perform spatio-temporal thinking, and now we are interested in understanding whether this finding can be affected by display type. Existing tasks in the literature, e.g., estimating the relative orientation of a boat to infer spatial performance in 2D [TGSP03], or using Tri-dimensional chess for 3D [SLU+96], are not appropriate for panoramic video imagery which includes both space and time reasoning in hybrid panoramic space. Similarly to previous experiments then, to quantitatively assess spatio-temporal understanding, we need to design tasks for participants to complete which can reveal insights into the way users perform spatio-temporal thinking.

Following from the discussion introduced in Section 6.3, and learning from previous experiments, it is clear that common actions while virtually exploring a place include looking for objects/actions in space and in time, following dynamic events within the place, and identifying when changes happen within specific times or areas of the place. These considerations are supported by the applications for which Vidicontexts was originally developed – i.e. surveillance, virtual tourism and telepresence applications – and which have been discussed in Section 6.3. As such, we decided to employ the counting and tracking tasks designed for the Vidicontexts experiment (see Section 6.3), as they offer two reliable metrics which a) mimic common tasks performed when exploring panoramic imagery, and b) are not dependent on the display device used. Therefore, we split our experiment over two tasks, one focussed on counting people located in a particular area, and one involving tracking people that crosses a certain region of interest, and

we analysed three display modes: a flat desktop display, and egocentric tablet display and an immersive HMD device. We also decided to focus our analysis on similar parameters as the one used in previous experiments.

Hypotheses In both tasks, we measured the completion time and accuracy expressed as errors in the people counts, and we collected the results of usability and tasks related questionnaires. Immersive displays such as HMDs might be more suitable to display panoramic representations as they are egocentric, allowing for natural navigation of the environment with head rotation. Tablet devices are less immersive as only a portion of the view is taken up by the virtual window, but tablet use can still be egocentric by rotating the device in space. Desktop displays are exocentric, and so immersion is reduced further. With these premises, and following the results of previous studies which showed that immersion might increase accuracy [SLU+96, PSP93, TGSP03, PCS+00], we expect accuracy to vary with the level of immersion of the display, and so we expect the HMD display to be most accurate, the tablet display to be less accurate than the HMD, and for the desktop display to be least accurate.

While input devices differ across displays, we do not expect completion time to vary significantly. As previously discussed, the major workload is in spatio-temporal reasoning and not on interface manipulation. We expect the three conditions to obtain SUS scores relative to their familiarity, with the desktop display obtaining the best score followed, in turn, by the tablet and then the HMD. For the task questionnaire, we expect all three conditions to indicate that the interface was suitable for the task, but we expect exocentric views to be preferred over egocentric views as readability is higher, as per [MSD+12].

7.3.1 Method

Participants

30 unpaid participants from the staff and student population at our university performed both tasks using one of the three displays each for a between-subjects design for the display type independent condition, and a within-subjects design for the task. While we did not filter the study population for handedness and eyesight, we ensured gender balance was respected. Additionally, the participants were randomly assigned one of the three systems, and there was no mention of the overarching goal of the study. Participants were recruited via e-mails and other forms or messaging.

Design

The two tasks adopted in the study intended to explore whether a certain display type affects the accuracy with which participants can correctly obtain a spatial and temporal understanding of a video collection. Both tasks involved counting and tracking objects in space and time. Hence, to assess performance, in both tasks we measured the completion time and accuracy expressed as errors in people counts. Further, following the experiment, participants completed the standard System Usability Scale (SUS) questionnaire [Bro96], and they were asked to answer height task-related questions (see Appendix F and Figure 7.7 for a list of questions). Their impression on the experiment was also recorded.

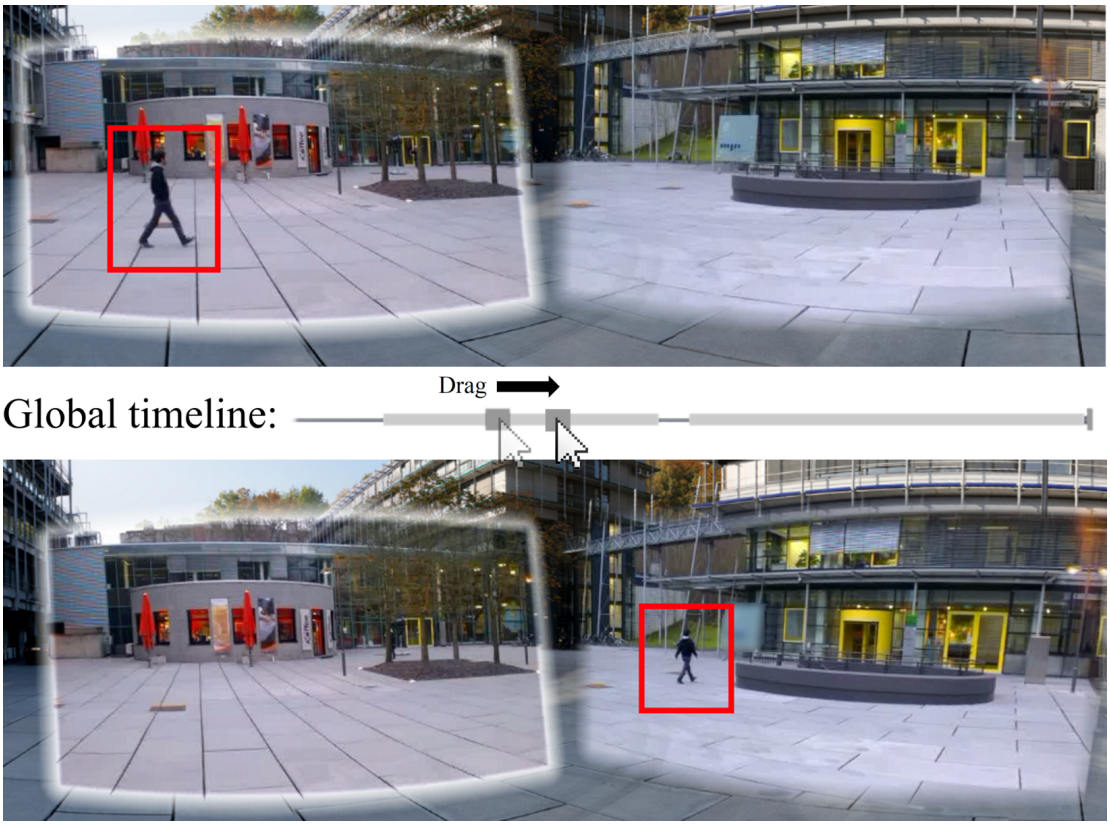


Figure 7.3: *A demonstration of the* tracking *task, where the number of unique people passing between two buildings must be counted.*

Procedure

As already mentioned, our study is split over two tasks and three video modes. As the number of videos in each task is reasonably small and as human action in video is memorable, there is a large potential for participants to learn the content if we conducted a within-subjects experiment. Instead, we conducted a between-subjects design to prevent this effect.

The first task — the *tracking task* — asks participants to review 6 videos and count the number of different people who cross between two buildings in a scene (Figure 7.3). Here, the videos never fully track a person and do not overlap, so multiple synchronous videos must be analysed to obtain the correct result; however, the task can be entirely completed to a high accuracy by manipulating the global timeline and focusing attention on a specific spatial region in the panorama. The dataset was collected in a university courtyard. Videos differ in length (2.30 – 4.00 minutes) but are time sequential, and they cover 125° horizontally within the environment. A participant could potentially make 12 erroneous counts (manually verified).

The second task — the *counting task* — required users to browse 20 videos and identify the number of different people who sit on a set of benches within a neo-classical quadrangle (Figure 7.4). Videos differ in length (0.20 – 1.10 minutes), are not time sequential, and cover the entire horizontal 360° extent of the environment. This task requires users to spatio-temporally reason much more than the previous task as the same people appear in multiple video foci at different times, with some people sitting only

Figure 7.4: *Task interface, here showing the* counting *task. Participants need to identify and count unique people who sit on the benches positioned below the columns. While this might seem simple, this is a complex spatio-temporal reasoning task as people appear in multiple video foci at very different times. This task requires tracking the entrance and exit of persons across the scene space and time to verify their identity.*

near the areas of interest or standing in front of the benches. Participants must focus on parts of the panorama which are farther apart to track the entrances and exits of people and verify their whereabouts and identities. A participant could potentially make 20 erroneous counts (manually verified).

Before to start the experiment, each participant was given a detailed description of the interface features, and as much time as they wished to familiarise before the tasks. Participants could use all features of each system. Then, each task was conducted in series, in random order, and under no time limit, with a briefing beforehand to explain the task. Following both tasks, the participant completed two questionnaires and their impressions, if any, were recorded.

7.4 Study Results

The primary dependent measures of interest used for both tasks were the accuracy expressed as errors in the people counts and the time taken to complete the task. Table 7.2 shows an overview of the results obtained in the two tasks. Initially, for statistical analysis a 3×2 (display $\times$ task) mixed Analysis of Variance (ANOVA) was computed using SPSS [IBM09] to analyse each of the dependent variables. Display type was a between-subject factor, while task was a within-subject factor. A significance level of .05 ($\alpha = 0.05$) was used for judging the significance of effects. No samples were removed from our analysis, as all the participants successfully completed their tasks.

7.4.1 Accuracy

The counting error results shall be considered first. Figure 7.5 presents a summary of the accuracy results. It shows mean and standard deviation for the number of errors committed, as well as the statistical

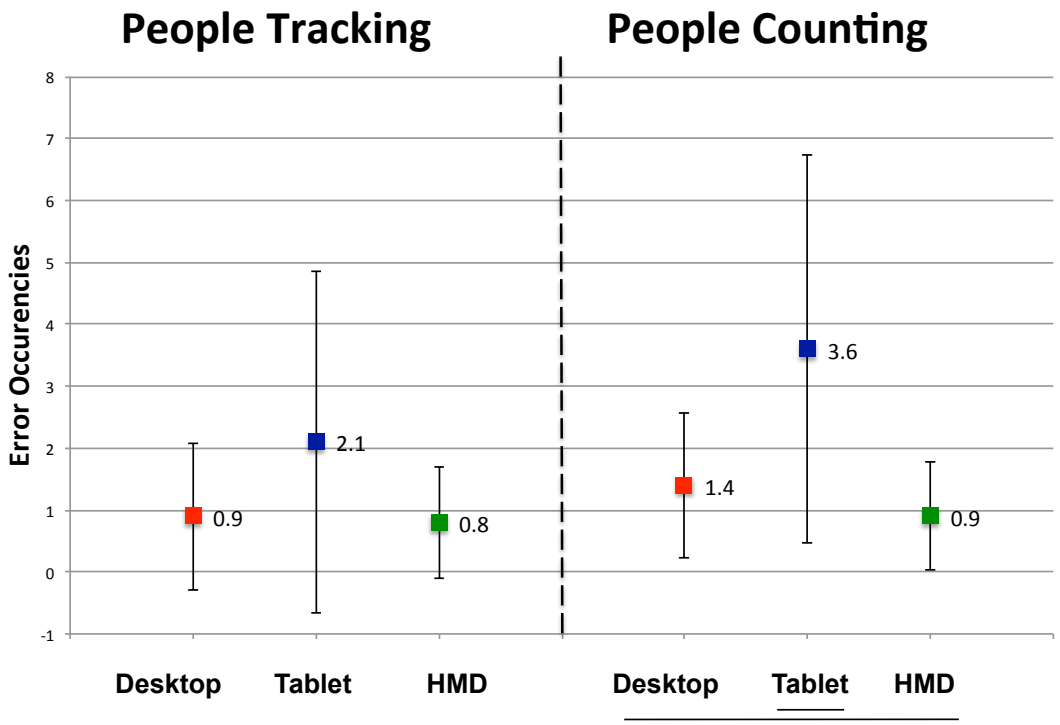


Figure 7.5: *Mean counting errors and standard deviation for each display type and task. When a main effect of display is found, conditions jointly underlined are statistically similar (or a trend is found).*

Condition	Tracking Task			Counting Task		
	Error	Time (sec.)	Normalised Time	Error	Time (sec.)	Normalised Time
Desktop	0.9	373.36	0.332	1.4	469.12	0.521
Tablet	2.1	382.50	0.340	3.6	616.00	0.684
HMD	0.8	377.40	0.336	0.9	458.80	0.509

Table 7.2: *Tasks results. Normalised time is per frame over all video foci. Participants were approximately twice as fast per frame of video at the tracking task as it involved constantly comparing two video foci at once within the panorama.*

significance, for each task and condition combination.

For the *tracking task*, participants had fewer errors for the HMD ($M = 0.8$, $SD = 0.91$) and desktop cases ($M = 0.9$, $SD = 1.19$) than for the tablet ($M = 2.1$, $SD = 2.76$). Similarly, for the *counting task*, there were fewer errors for the HMD ($M = 0.9$, $SD = 0.87$) and desktop cases ($M = 1.4$, $SD = 1.17$) than for the tablet ($M = 3.6$, $SD = 3.13$). Results of statistical analysis found a main effect of display type on accuracy ($F_{(2,27)} = 7.208$, $p = 0.003$). However, no main effect of task on accuracy was reveal ($F_{(1,27)} = 1.805$, $p = 0.190$). The interaction between display and task was also not significant ($F_{(2,27)} = 0.638$, $p = 0.536$).

To further unpack the effect of the between-subject factor (i.e., display type) at each level of task, we computed two ANOVA (one per task) with the display type used as the single factor and counting error as the dependent variable, with post-hoc Tukey tests for pair-wise significance tests. For the *tracking task*, a non-significant main effect of display type was found for accuracy ($F_{(2,27)} = 1.581$, $p = 0.224$). On the

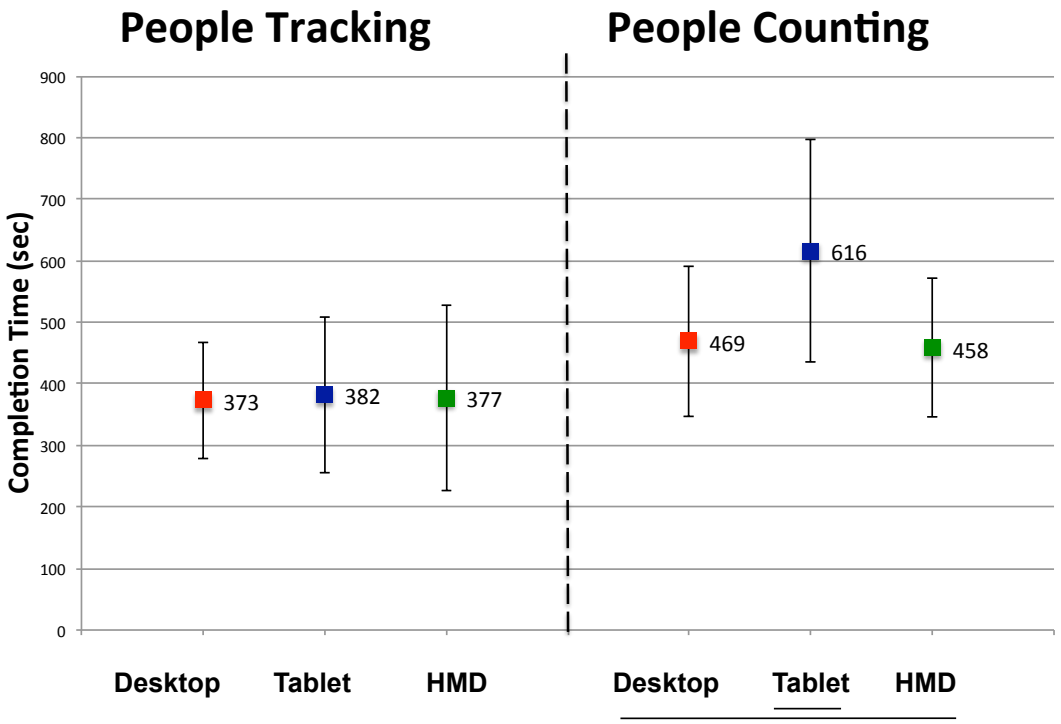


Figure 7.6: *Mean completion time and standard deviation for each display type and task. When a main effect of display is found, conditions jointly underlined are statistically similar (or a trend is found).*

contrary, for the *counting task*, display type was a significant main effect for accuracy ($F_{(2,27)} = 5.173$, $p = 0.013$). Post-hoc analysis revealed significant differences between HMD and tablet ($p = 0.015$), a trend for significance between desktop and tablet ($p = 0.052$), but no significant difference between desktop and HMD ($p = 0.842$) and.

7.4.2 Time to Complete

We will now analyse the dependent variable completion time. Figure 7.6 offers an overview of the completion time results. It shows mean and standard deviation for the time to complete, as well as the statistical significance, for each task and condition combination.

For the *tracking task*, the desktop display obtained the lowest mean time ($M = 373.36$ sec., $SD = 94.31$ sec.), followed by the HMD ($M = 377.4$ sec., $SD = 150.35$ sec.) and the tablet ($M = 382.5$ sec., $SD = 126.91$ sec.). Regarding the *counting task*, the HMD case obtained the lowest mean time ($M = 458.8$ sec., $SD = 112.41$ sec.), followed by the desktop ($M = 469.12$ sec., $SD = 121.85$ sec.) and the tablet ($M = 616$ sec., $SD = 180.36$ sec.). Analysing the statistical results, no main effect of display type on accuracy can be seen ($F_{(2,27)} = 1.933$, $p = 0.164$), but a main effect of task on accuracy was reveal ($F_{(1,27)} = 20.055$, $p < 0.001$). The interaction between display and task was also not significant ($F_{(2,27)} = 2.516$, $p = 0.1$).

To further analysing the effect of the between-subject factor (i.e., display type) at each level of task, we computed two ANOVA (one per task) with the display type used as the single factor and time to complete as the dependent variable, with post-hoc Tukey tests for pair-wise significance tests. For the *tracking task*, a non-significant main effect of display type was found for completion time ($F_{(2,27)} = 0.13$,

Task-related question	Desktop	Tablet	HMD
Q1: Easy to complete tasks	4.0	4.5	3.66
Q2: Understood video orientation in space	4.7	3.9	3.8
Q3: Understood relative video position	4.4	4.2	3.6
Q4: Understood space-time video overlap	4.3	4.1	4.0
Q5: Understood temporal order of videos	3.4	3.3	3.3
Q6: Environment representation confused	3.3	3.1	2.5
Q7: System has enough functions for tasks	4.4	4.1	4.0
Q8: #videos made remembering things hard	2.4	1.6	1.9
Overall mean	3.86	3.6	3.34

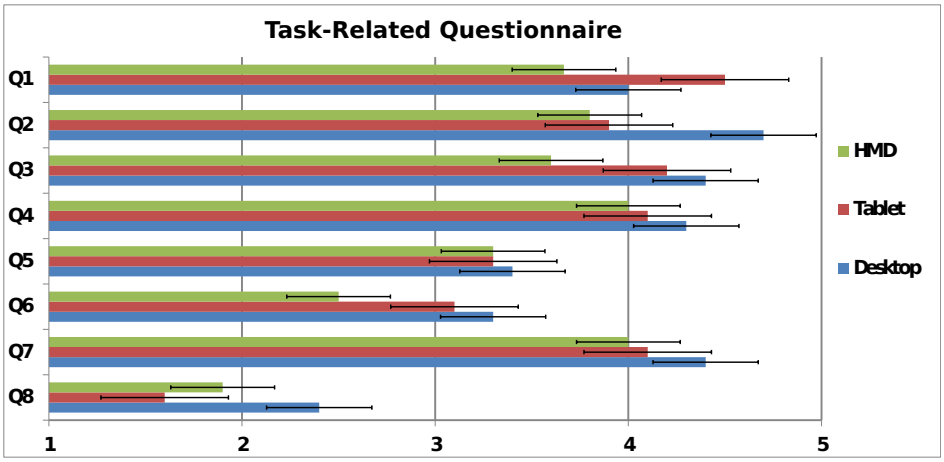


Figure 7.7: *Mean and variance plot for the task-related questionnaire. We assume Likert ordinal data was fairly interpreted as an interval scale, with text labels ranging from* strongly disagree *to* strongly agree. *The scale for negative questions was reversed for mean computation.*

$p = 0.987$). However, for the *counting task*, display type emerged as a significant main effect for time to complete ($F_{(2,27)} = 3.865$, $p = 0.033$). Post-hoc analysis revealed a significant difference between HMD and tablet ($p = 0.049$), confirming that the HMD allows user to perform their task faster than tablet users. There was also a statistical trend showing that desktop were more performant than tablets ($p = 0.07$). However, no significant difference between desktop and HMD ($p = 0.985$) was found.

7.4.3 Questionnaires

For the system usability scale, both desktop and tablet cases scored above average ($SUS = 77.5$ and $SUS = 76.5$ respectively), tailed by the HMD case ($SUS = 68$). Following the SUS classification technique of Lewis et al. [LS09] (letter-grade ranks varying from A to F), the desktop and tablet cases are Rank B systems, while the HMD version is a Rank C system. Rank A systems have many promoters, who will definitely use and recommend the product, while rank B systems have a fair number of promoters, who are likely to use and promote the product. All other ranks will only have detractors.

For the task-related questionnaire, across all questions, there were no significant differences (Figure 7.7). We can conclude that participants felt capable of completing the tasks on all three displays, that all three provided a good sense of orientation and allowed the relative position of videos to be understood,

that all three allowed spatio-temporal reasoning and did not induce spatio-temporal confusion.

7.4.4 Observations

To complete the tasks, 80% of the population assigned to the desktop condition used equirectangular projection. This finding shows that participants thought the 360°-at-once projection conferred more spatio-temporal information than the perspective projection, as expected. Regarding tablet users task strategy, 90% of the population preferred to use touch-based rotation rather than orientation rotation navigation. Almost all users first attempted to use the orientation sensor-based navigation, but then switched to touch-based navigation. We discuss the implication of this observation in the next section. No one reported eye strain or nausea for the HMD case. HMD users regularly used zoom controls, in contrast to desktop and tablet users who used zoom controls very rarely. This can be explained by the low resolution of the HMD display in comparison with the desktop and tablet displays. Across all conditions, no one struggled to finish the tasks.

7.5 Discussion

7.5.1 Display Effects

For the tracking task, display type was not found to be a significant factor for either completion time or accuracy. We conclude that the task was sufficiently simple that the display type did not make a difference and all three displays are suitable for simple tasks. This shows the importance of using a complex task when assessing spatio-temporal reasoning (supporting [SLU+96], contrasting [TGSP03]). However, the smaller display on the tablet may be a cause of increased time and errors for some people, as the variance for the tablet condition is higher than for both desktop and HMD conditions.

For the counting task, the tablet took significantly longer than the HMD and, even though not significant, there is a statistical trend suggesting that it takes longer than the desktop too. Similarly, the tablet is significantly less accurate than the HMD and, even though not significant, seems less accurate than the desktop too. The distributions in both time and accuracy show much larger variance in the tablet case, and this follows the trend in the other task.

Our hypothesis is not true in our experiment, as task accuracy does not allow us to conclude that display immersion can be considered a significant factor for panoramic imagery. Effectively, users were able to achieve equivalent levels of accuracy in both non-immersive (desktop) and immersive (HMD) displays. This suggests that the potentially positive performance effect of immersion in 3D environments does not necessarily extend to panoramas for either our tracking or counting tasks. Further, results from both tasks indicate that egocentric immersive views can be as performant as exocentric non-immersive views in completion time and accuracy.

7.5.2 Tablet

The tablet condition appears worse than the desktop, and was significantly worse than the HMD in the complex task. We suggest that the smaller tablet display, even though it is high resolution, negatively affected spatio-temporal reasoning. From observing participants solving strategies, we did not notice

participants zooming or bringing the tablet closer to see more detail. Further, after an initial period of using orientation sensor rotation, nearly all participants switched to touch rotation. This does not explain the longer completion times as, for the simple task, times are comparable across all devices and task order was random. When asking participants to explain why they switched to touch navigation to complete the tasks, participants cited: 1) that camera movement was too tied to device movement, making navigation confusing; 2) that holding the tablet and interacting with the screen was too cumbersome (cf. [WHM12]), and 3) that the device was too heavy to hold in this way for long periods of time.

One might think that tablet resolution was a factor. For 1080p at 25 degrees field of view, each pixel on the tablet equals 0.6 arcminutes of view, in contrast with human eye acuity of approximately 1.2 arcminutes. As the tablet is mobile, this extra resolution could be viewed by simply moving the tablet closer. However, in general, this is a moot point and does not hinder performance, because the focus areas of the task — the people in scenes — with no zoom, are typically 10-30 pixels wide, and 250–2250 pixels in area.

Interestingly, the task questionnaire suggests participants felt the tablet was just as capable as the desktop, and the SUS scores suggest participants felt it was just as usable, too. This does not align with real task performance, which was reduced for the complex task. We suggest this is a familiarization issue, as participants were comfortable in general with tablets. This 'false sense of security' is potentially dangerous if tablets were to be used for critical panoramic review tasks, such as the ones to be performed in surveillance, remote assistance or panoramic telepresence applications.

7.5.3 HMD

The HMD performed similarly to the desktop, and significantly better than the tablet, However, the questionnaires scores suggest that users found it less capable, and the SUS scores suggest that users found it less usable. While one might think that the HMD was rated as less capable or usable for human reasons (eye strain, tiredness, nausea, general uncomfort), our participants reported no such issues. Instead, we suggest that this is a familiarization issue again, but now the reverse effect where the novelty of the device induces caution in qualitative assessment. However, given the equivalent performance to the desktop case, there is no reason to suggest that the HMD interface is a compromise for our tasks. Again, one might think that resolution would be an issue, as the HMD has 10x lower perceived resolution (with 12 arcminutes, in contrast to the desktop with 1.2 arcminutes). However, with simple zoom controls, the two display types performed similarly. This suggests that the tasks performance does not differ simply from a change in resolution perception.

The lack of benefit from using an HMD over a desktop in our experiment, expected from the immersion suggestions from virtual environment works [SLU+96], is unlikely to be attributed to the difference between rendered and photographed views, as Willemsen and Gooch suggest [WG02]. Instead, we suggest this parity instead comes from the added warped field of view provided by the exocentric equirectangular projection on the desktop.

7.5.4 Design Implications

We wish to discuss the potential generalization of our results. Many works in this field (and others) provide evidence for more general conclusions from a single experiment [SLU+96, TGSP03], which helps form a body of evidence within the literature for the general conclusion. We have seen effects in specific tasks that are limited to panoramic video imagery; however, for corroboration, the existing work concerning devices and panoramic imagery is limited. As explained in the introduction, this is imagery used daily by thousands of people, and so from our experience, task-based study, and questionnaires, we suggest implications of our study for these and other applications with similar video+context components. However, we caution the reader from drawing implications beyond our scope, we anticipate a continued scientific discussion on the effects of panoramic imagery. Additionally, given the nature of our results, we invite the reader to consider the following discussion as a list of suggestions:

- During our trials, participants preferred exocentric equirectangular projections over perspective projections on desktops, and this confirms previous study in literature [MSD+12]. This projection type seems to be an appropriate default for panoramic spatio-temporal reasoning task systems.
- Most HMD users frequently used zoom controls while performing both tasks, and our design choice to provide such controls was praised by several users when their impressions were recorded. The same did not happen for the other displays. Hence, we believe that it is important to provide zoom controls to overcome the comparatively low-resolution of some HMDs.
- Participants preferred touch rotation controls over arm-based orientation controls for tablets, as it is difficult to both orient and manipulate on-screen elements. This suggests that the ability to pause orientation control is necessary. However, even with this option, for our reasoning tasks participants reverted permanently to touch rotation. In-situ browsing and augmented reality situations may provoke a different response, given the nature of the AR interaction metaphor for which holding the portable device in front of the user eyes is paramount for successful augmentation. However, we believe that arm-based orientation controls are not recommended for tasks requiring long periods of concentration as they are tiring, and this is in line with previous findings in literature [BBL93, TFK+02].

7.5.5 Conclusion and Limitations

The experimental work described in this chapter reveals insights on the effect of display type on spatio-temporal reasoning. In particular, results showed that the three displays investigated are equally performant for our simple spatio-temporal reasoning tasks, but that for more complex tasks, such as the counting task, our tablet interface was less accurate and took more time than the HMD and desktop displays. Interestingly, participants perceived the tablet to be just as capable and equally usable as the desktop, even though task performance was worse. Contrarily, participants perceived the HMD to be less capable and less usable than the desktop, even though task performance was the same.

In Chapter 1, we hypothesised that the level of immersion of a display type can be a significant factor on users spatio-temporal thinking, affecting the eventual beneficial properties offered by the

video+context representation (i.e. H5). Even if contradicting our initial hypothesis, the results obtained here have implications on both the video+focus representation and for designing panoramic imagery systems. Firstly, the positive results showed in previous chapters can be extended to immersive displays, as no significant differences in performances can be found across display types. This makes the video+focus representation suitable for a vast range of applications and displays, a property particularly beneficial for asymmetric systems such as BEAMING. Supported by these results, while developing applications that leverage the videos+focus paradigm, we can be sufficiently confident that similar user performance on spatio-temporal reasoning can be achieved on both immersive and non-immersive displays.

The fact that tablet's users performed poorly, but perceived their performance positively, is an important factor which has implications for panoramic imagery interfaces. While designing future panoramic applications for tablet, researcher and designers should take this factor into account, especially for critical applications. Conversely, HMD users underestimated their performance. This suggests that, while designing applications for this type of display, this negative bias should be taken into account, and perhaps longer familiarisation times should be given to the user to minimise this effect.

In general we can conclude that immersion does not seem to be a significant factor while interacting with videos in panoramic context. However, especially for demanding tasks, such as the counting task, we suggest to adopt large FoV displays to minimise the confusing effect that could be introduced by the rich visual stimuli given by the proposed representation.

So far we have focussed the discussion on the results and implications of our study. However, it is important to note that, while the experimental design and consequent results helped us answering one of our research questions, other routes could have been explored during our investigation. We motivated our choice of tasks by grounding them in real-world usage of our system (see Section 7.3), and we were inspired by other experiments conducted on the Vidicontexts system for designing our investigation. However, other tasks could have been chosen, and in general all the alternatives presented in Section 6.5.3 would have applied to the system adaptations presented here. One important aspects to consider though, is that during our exploration we were interested in assessing the effect of display types on a particular representation rather than how well a system performed. To this end, it is important that the tasks selected mimic fundamental actions performed when exploring and reasoning about panoramic imagery, acting in this way as a proxy for other possible applications. Thus, selecting tasks that resemble one particular usage more than others (e.g., a surveillance-based task) could have produced too specific and less generalisable results.

If we consider the displays involved in the study, more and different displays could have been introduced in the evaluation. Especially large projection displays and CAVE systems would have added an interesting point of comparison. However, in this iteration of the study we decided to focus our comparison on displays that are easily accessible and represent real-world setup for our system. Nevertheless, we reserve to extend the comparison to other type of displays in future work.

In the current experimental design the usage of the tablet device was restricted to an office space which largely differed from the contexts depicted on screen. However, a tablet device offers a power-

ful instrument for in-situ browsing and augmented-reality experiences which can greatly improve user performances over remote explorations [RHQ14]. Hence, another interesting point of comparison could have been in-situ exploration of the media collection, in which a user would have performed the given tasks while being physically collocated in the locations were the contexts was firstly recorded. This "collocated" conditions could have been either used in conjunction with the other display types selected in the original design, or in an investigatory experiment in which, similarly to [RHQ14], both collocated and non-collocated tablet conditions could have been compared to select the best option.

Finally, it is important to discuss the external validity of our study. During our investigation we have seen effects in specific tasks that are limited to panoramic video imagery; however, for corroboration, the existing work concerning devices and panoramic imagery is limited, and therefore we caution the reader from generalising our result beyond our scope. However, as we have seen effects that go against similar studies in related discipline (i.e. Virtual Environments, albeit on different visual stimuli), we anticipate a continued scientific discussion on the effects of panoramic imagery. To this end we believe that our results can be beneficial for future researchers which will investigate other aspects of this topics, or for developers wanting to build effective videos+context interfaces.

Another concern with the generalizability of the findings of our study is that the tracking task was artificially simple. In the study participants were required to track different person over videos which could be easily spatially and temporally compared using our interface. Additionally, users knew a priori the areas of interest. This however is not the case for real-world scenarios, where the area of focus is not known in advance. Perhaps, a different question to ask in future studies is whether users are able to identify areas of interests prior tracking, and if they are, whether this is influenced by the display, the interface, or a combination of the two. It would also be interesting to explore whether different versions of the Vidicontexts interface, with increasingly reduced features, can affect user performances, and whether there is a dual effect of system features and display type.

7.6 Chapter Summary

This chapter presented an investigation on immersive display effect on panoramic spatio-temporal reasoning tasks. To create one simple and one complex reasoning task, we exploited the novel panoramic video foci idea presented in Chapter 6 and created an adaptive multi-display interface. We conducted a user study with desktop, tablet, and HMD displays covering exocentric and egocentric modes.

The chapter has been structured as follows. In Section 7.1 we introduced the motivation of the user study, briefly presenting similar works conducted within the virtual environments research community, and discussing the usefulness of the study on the overarching research theme of this thesis. We then described the extensions made to Vidicontexts to fit two additional display types, a tablet display and an HMD, briefly presenting the two interfaces. The experimental design was then introduced, followed by a report on the results collected during the study. Finally, we discussed the implication of the study results with respect to display type, application design and, most importantly, the video+focus representation.

In our investigation we discovered that desktop and HMD devices perform comparably, even if users feel the HMD is less capable and less usable. We find that tablet displays are significantly less per-

formant in our complex spatio-temporal reasoning task, even though participants found them as capable and usable as a desktop. These results form implications for panoramic imagery interfaces for spatio-temporal reasoning tasks, ad confirm that the results collected in previous chapter can be extended to a variety of display types. This last factor complements the results presented so far, giving interesting points of discussion for the overarching research theme of this thesis. The next chapter then will summarise the findings obtained in the three user studies presented in this thesis, relating back to the research overarching goal and questions to obtain a complete analysis of the videos in context representation.

Chapter 8

Discussion

The aim of argument, or of discussion, should not be victory, but progress.

Joseph Joubert

The quality and pervasiveness of cameras on mobile devices continues to increase. Most new laptops have a built-in camera, and most new smartphones and tablet-style devices have both front- and rear-mounted cameras. Rear-mounted cameras on mobile devices aim to replace or supplement the use of a point-and-shoot camera, while front-mounted and laptop cameras are often used for face-to-face video conferencing. As a consequence, panoramic images and video are now common, with the world quickly being mapped at street level by companies and tourists alike.

While the abundance of cameras and panoramic imagery can be exploited for applications where spatial understanding of the scene is critical, such as surveillance and collaborative telepresence applications, combining videos and panoramic imagery in a single representation presents many challenges to providing useful interfaces to the content.

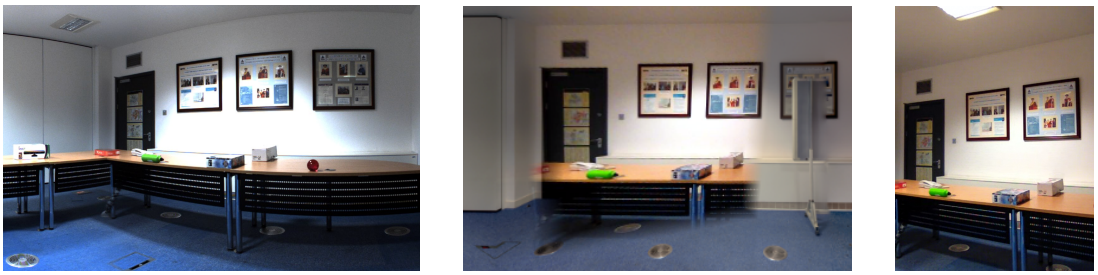
On a spectrum between 3D virtual environments and 2D images, panoramas lie somewhere in between – a 360° panorama can surround a user, but the scene has only spherical geometry and is effectively flat. The user cannot move from their point of view and so does not receive parallax cues. However, when augmented with live video insets, panoramic imagery can convey spatial and temporal information of a remote scene which can greatly benefit users. We call this novel visual representation *videos in panoramic context* or in short *videos+focus*, and we study its properties with controlled experiments whose results are reported in this thesis.

The aforementioned user studies are presented in Chapters 5–7, and each chapter focuses on a particular property of the videos in panoramic context representation. Chapter 5 investigates the suitability of a single video in panoramic context for a collaborative telepresence scenario. Chapter 6 explores the effect of video-collection in context on user spatial and temporal understanding of a remote scene and the dynamics within. Chapter 7, finally, studies the effect of display type on users interfacing with multiple videos in panoramic context.

The overarching goal of this experimental work is to explore how videos in context may be employed to convey spatial and temporal information of a remote location, and how well this representation

can replace more sophisticated visual descriptions. The rest of this chapter then summarises the experimental work presented in this thesis, relating back to the thesis overarching goal and to the research questions and contributions established in Chapter 1.

8.1 Videos in Context for Telecommunication



(a) *Panoramic.* **(b)** *PanoInserts.* **(c)** *Webcam.*

Figure 8.1: *The three telecommunication systems used in the study presented in Chapter 5. Panoramic and PanoInserts videos are cropped for illustration purposes, but both use the same equirectangular projection.*

With the user study presented in Chapter 5, we investigated the suitability of a single video in panoramic context for collaborative telepresence scenario. Results of the study showed that our proposed representation, demonstrated with a system which we developed and call *PanoInserts*, can be directly compared to fully panoramic videos and outperforms standard webcam style video-chat in tasks requiring a high level of spatial reasoning.

Implications of this result are manifold. The video in context representation can replace the more sophisticated and expensive fully-panoramic video without loss of performances. Fully panoramic video or large FoV cameras are often used in highly-developed video conferencing systems, such as Cisco *TelePresence* [Cis06], requiring however intrusive and limiting technical interventions. As a consequence, since our representation can be quickly acquired exclusively using mobile devices, our solutions enables surround, portable video conferencing featuring high-end system's spatiality while preserving ubiquitous webcam-style video chat portability. Indeed, employing similar hardware as the more common portable webcam-based systems, our solution replaces webcam style video-chat, improving the communication experience with few simple additions.

When relating these findings with the main research question of this thesis, this first study suggests that video in panoramic context can indeed be used to describe remote location, successfully conveying its spatial properties. Users can intuitively understand and act upon the proposed representation, achieving a good level of spatiality while perceiving a clear visual stimuli. Evidence of this are given by both the experimental tasks results and users strategies in conducting them. Users of both our system and the panoramic video system successfully infer not only general spatial information on the remote room, but also detailed spatial information on the objects within. Relative positions of the objects, as well as position with respect to other objects or room's landmarks, are successfully recovered and often used. However this depends on the spatial richness of the stimuli, and does not hold if the spatial nature of

the perspective view is impoverished as in the webcam condition. This suggests that systems using the video in panoramic context representation, similarly to high-end fully panoramic video systems, support spatiality properties, such as movement and distance, shared frame of reference and containment [BGR+98], which are essential factors to improve remote interactions [HRBC06, VWS02, SNO+12].

While revealing fundamental implications for this thesis, this first study has some limitations that prevents us to generalise its findings. First, the representation investigated here is limited to a single video in context. While this allowed a fair comparison to an existing webcam system, investigating the more general case of multiple videos in context would give us information on the full potential of the representation. Second, we only explored scenarios where equirectangular projections were used for the panorama. While this is a common strategy (see Mulloni et al. [MSD+12]), other projections (e.g., perspective projection) are available and could affect users perception differently. Therefore, to extend the findings of this first experiment, we conducted a second user study which we discuss next.

8.2 Videos in Context for Spatio-Temporal Browsing

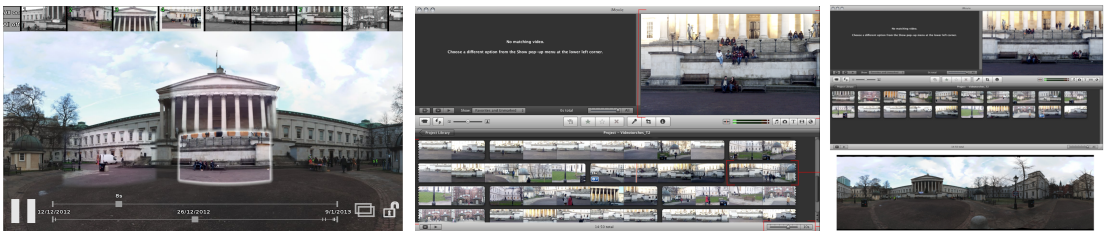


Figure 8.2: *The three video collection browsing interfaces used in the study presented in Chapter 6.* Left: *Vidicontexts;* Center: *iMovie;* Right: *iMovie with panoramic reference.*

With the user study presented in Chapter 6, we investigated the suitability of multiple videos in panoramic context for spatial and temporal coherent browsing. Results of the study showed that our proposed representation, demonstrated with a system which we developed and call *Vidicontexts*, performs better than typical video browsing tools in tasks that require to infer spatial and temporal properties of a remote space, providing significant benefits to accuracy and time taken in such localization tasks.

Implications of these results are manifold. Through our experiment, we can establish that contextualising video collections with panoramic imagery is not only beneficial for users' spatial reasoning, but it also improves their temporal understanding of the video collection. With our system validation then, we found that the proposed representation is positively perceived by users, which can intuitively understand it and act upon it. This finding is not trivial, as our representation presents a richer visual stimuli than existing (and largely diffused) video browsing tools.

Analysing users' strategy in completing the various tasks, we note that environment's landmarks, which are a fundamental addition of the context, are often exploited by the users of our system to infer spatial relationships between videos of the collections and between people and objects within the videos. This observation suggests that the spatiality properties of the video in context representation found in PanoInserts extend to the richer visual representation presented here. Supported by these results and

previous experiment, we can conclude that videos in panoramic contexts help users building spatial and temporal maps of remote places, thus enhancing the level of spatiality supported by the system. On a practical level, this means that our representation can reduce spatio-temporal task's cognitive load by automatically building spatial and temporal links between videos of a collection. The videos+focus representation implicitly offers to the user a spatio-temporal map which can be used to navigate both the collection and the context. In line with this result, we note that the majority of the population assigned to our system used equirectangular projection to complete the tasks. This finding shows that participants thought the 360°-at-once projection conferred more spatio-temporal information than the perspective projection, as expected [MSD+12].

The results of the study, then, confirm the ones presented in previous section and extend them to the general case of many videos in context. It is clear that the videos in context paradigm offers a beneficial representation for both video-conferencing systems and video browsing. Supported by these result, we can conclude that, as providing panoramic context is a special case of the general problem of aligning content to world model, the proposed representation offers a valid crutch to provide space-time exploration of remote environments and video collections.

With our studies we investigated visual properties of the videos+focus representation. To conclude our investigation on the representation though, we now move the focus on display types. As panoramic imagery fits a variety of display types that goes beyond flat screens, we want to know if the results presented so far can be extended to more immersive form of displays. Studying this has many implications, which we present and analyse in the following section.

8.3 Effect of Display Type on Videos in Context

Figure 8.3: *The three display types used in the user study presented in Chapter 7.* Left: *Flat display;* Center: *Tablet;* Right: *HMD.*

With the third and conclusive study of this thesis, presented in Chapter 7, we investigated the effect of display types on users interfacing with videos in panoramic context. Having established that the videos in context paradigm has clear benefits on users' spatio-temporal reasoning, we now focus on whether the display type used can be an affecting factor on users' performances. Specifically, we designed a user study to analyse three displays which sample interesting points within the immersive displays design space: 1) flat desktop displays; 2) mobile tablet displays with orientation tracking; and 3) HMDs with head tracking.

Results of the study showed that HMD users perform as well as desktop display users, and that both conditions outperform the tablet device, albeit this last finding is not always statistically signifi-

cant. Additionally, the study revealed interesting insights on how users approach the different devices. HMD users felt less capable in performing the given tasks, even if their performance was comparable to the more confident desktop users. Interestingly, we found this reversed for the tablet case, as tablet displays are less performant in our spatio-temporal reasoning tasks, even though participants found them as capable and usable as a desktop.

These results have implications on both the video+focus representation and for designing panoramic imagery systems. First, the experiment results does not allow to conclude that immersion is a significant factor while interacting with videos in panoramic context. This suggests that the positive results showed in the rest of the experimental work of this thesis can be extended to immersive displays, as no significant differences in performances can be found across display types. This makes the video+focus representation suitable for a vast range of applications and displays. Supported by these results, while developing applications that leverage the videos+focus paradigm, we can expect that similar performances on spatio-temporal reasoning tasks can be achieved on both immersive and non-immersive displays. However, researchers should take into account the negative self-assessment bias we found on tablet's users, and consider it while developing critical panoramic applications. Similarly, applications designed for egocentric HMDs should be aware of the fact that users need some familiarisation time before starting to feel confident with this relatively uncommon immersive display.

8.4 Conclusion

The results obtained from the experimental works presented in this thesis help us answering the research questions established in Chapter 1. The overarching goal of this experimental work was to explore how videos in context may be employed to convey spatial and temporal information describing a remote location, and how well this representation can replace more sophisticated visual descriptions. We presented two different interfaces to single and multiple videos in panoramic context, and evaluated the visual properties of the representation with three user studies. In line with our initial hypothesis, we established that videos in panoramic context are a suitable alternative to more sophistical visual representations, such as fully panoramic videos or 3D models, to improve users' spatial and temporal thinking. Additionally, we established that the rich visual stimuli provided by the video+focus paradigm can be easily understood and acted upon.

The benefits of the proposed representation can be found in the high level of spatiality conveyed by it and, whenever possible, in the fact that it improves temporal thinking. A crucial aspect of these properties is that they are intrinsics to the representation, and as such do not depend on the display used. This allows us to conclude that the video+focus representation can benefit a variety of applications and displays. Chapter 1 introduced three central research questions that are components of the overarching goal of this experimental work and which we fully addressed during our studies. Specifically:

1. *Can spatially localised video be used to increase the spatial information transmitted during video mediated communication, and does this improve quality of communication between users and their spatial thinking?*

Yes. We demonstrated that users of PanoInserts, our mobile videconferencing tool, can successfully infer spatial information of remote locations and objects within it, obtaining better performances than standard video-chat systems' users. Additionally, we found evidences that users of our proposed representation obtain spatial cues from a combination of the panoramic context and the objects within the live video foci. The same can be found in fully-panoramic video systems, but it cannot be found in standard video-chat systems.

2. *Can videos in panoramic context be used to convey spatial and temporal information describing a remote place and the dynamics within, and does this improve users' performance in tasks that require spatial and temporal thinking?*

   Yes. We demonstrated this with a user study on our spatio-temporal video browsing interface, Vidicontexts. We found that users exploring remote locations through our interface can achieve a higher quality spatial and temporal reasoning than users of conventional video browser tools. During our studies we established that the proposed representation encodes spatio-temporal information which are otherwise impossible to convey with standard video browsing tools, and which users successfully understand and act upon. This results in a high level of spatiality conveyed by the visual representation, which in turn improves spatial awareness and eases users cognitive tasks' load.

3. *Measured by spatio-temporal thinking, is there an impact of display type on reasoning about events within videos in panoramic context?*

   No. We demonstrated this with a user study that explored three adaptations of Vidicontexts on three display types which sample interesting points within the immersive displays design space. We found out that immersion, unlike the case of 3D environments exploration, cannot be considered a significant factor in our spatio-temporal tasks, and while reasoning about events within videos in panoramic context. While this result is in contrast with our initial hypothesis, it allows us to extend the findings of Chapters 5 and 6 to a variety of display types, including egocentric immersive HMDs.

Similarly, in Chapter 1 we defined five hypotheses related to various aspects of the proposed representation, which we validated during our user studies. Specifically, we have confirmed that the videos in panoramic context representation is able to:

- **H1**: *build a spatial and temporal graph of several videos/cameras shown together through the employment of a common, panoramic context;*

  This hypothesis was confirmed in Chapter 6, where the system developed and the studies conducted showed how videos in panoramic context help users understanding spatial and temporal relationships within remote location and video-collections, by offering a readily-available spatio-temporal graphs of several videos or cameras.

- **H2**: *obtain a comprehensive depiction of a remote location through dynamic videos and static imagery, improving users' spatio-temporal thinking, and consequently being beneficial for the system spatiality;*

  This hypothesis was confirmed with the user studies presented in Chapters 5 and 6. The experimental results demonstrated how the proposed representation is beneficial for system's spatiality as it encodes spatial information that users are able to understand and act upon.

- **H3**: *being achieved in a small amount of time (from few minutes to an hour, depending on the number of video streams employed), and with minimal technical intervention, relying solely on available hardware;*

  This hypothesis was investigated and validated in Chapters 5 and 6. In there, the system developed and studied demonstrated how the the benefits of our representations come with little technical effort achievable with common devices.

- **H4**: *improve the sense of space and, when possible, time.*

  Similarly to H2, this hypothesis was confirmed in Chapters 6 and, partially, in Chapter 5. Through the studies presented in those chapters we demonstrated how contextualising large video-collection through a spatio-temporal index and with the aid of static panoramas can help user improve both spatial and temporal understanding of remote places.

However, one of our initial hypotheses was rejected by our study. Specifically, we could not confirm that:

- **H5**: *the level of immersion of a display type can be a significant factor on users spatio-temporal thinking, affecting the eventual beneficial properties offered by the video+context representation.*

  This hypothesis was rejected by the study results presented in Chpater 7. In there, we established that the level of immersion of a display type cannot be considered a significant factor on users spatio-temporal thinking, and consequently we cannot conclude that it affects the beneficial properties offered by the video+context representation.

To summarise, videos in panoramic context offer a valid solution to allow remote location exploration. As such, they are a valid alternative to more sophisticated, and often expensive and inconvenient, solutions to visually represent remote locations and the dynamics within. These findings are beneficial for a vast range of applications, such as virtual tourism, remote assistance or teleconferencing. Most importantly, some of the results of our studies are extremely relevant to BEAMING. Discussing this further, we can identify four main beneficial aspects of the video+focus representation with respect to BEAMING:

- BEAMING's technical intervention must be portable, self-calibrating, and dynamically configurable. In short, it should be as unobtrusive as possible so that it does not interfere with the locals'

behaviour. We demonstrated how the videos in panoramic context representation can be achieved through minimal technical interventions, including consumer mobile devices.

- ICVE systems such as BEAMING have to support spatiality properties, such as movement and distance, shared frame of reference and containment [BGR+98], to improve remote interactions [HRBC06, VWS02, SNO+12]. We demonstrated how the video+focus representation supports such properties, enabling a good level of spatiality.

- A technical requirements of BEAMING is that its sessions must be re-playable. The videos in context representation naturally fulfils this requirement, and the fact the representation benefits users' temporal reasoning is indeed a positive aspect which can be beneficial to the users during session replay.

Chapter 9

Conclusions

Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information.

David Marr

This thesis has investigated the suitability of videos in context for telecommunication and spatio-temporal browsing of video-collections. Telecommunication is increasingly being carried out in multi-user VEs and VMC systems, in which users interact within a shared virtual space. Millions of people work, play, and socialise for large amounts of time in online virtual worlds such as Linden Research's *Second Life* [Lin03] or telepresence systems such as Cisco's *Telepresence* system [Cis06]. However, enabling remote locations to be used within such systems is usually a tedious process that requires either manually modelling of the remote environment or the employment of specific hardware. At the same time, the increase in quality and pervasiveness of portable devices' cameras increased the amount of visual information present online, with the world quickly being mapped at street level by companies and tourists alike. This resulted in a large amount of capturing devices and, consequently, available visual information which can be employed to a) capturing and transmitting video from virtually anywhere; b) reconstruct environments from unstructured video collections; and c) spatially and temporally organise videos.

Inspired by focus+context systems, in which a subset of information is shown in full detail within a wider context of surrounding lower-density detail [CKB09, BGBS02], we propose a visual representation in which videos are aligned to a panoramic context to create a dynamic reconstruction of remote environments to be used for both VMC and browsing applications. We identify panoramas as a valid solution to the challenging problem of aligning video content to the real world for a variety of reasons. First, we note that panoramic imagery and videos are now common, with users being able to capture them using both dedicated hardware and consumer portable devices. Second, on a spectrum between 3D virtual environments and 2D images, panoramas lie somewhere in between – a 360° panorama can surround a user, but the scene has only spherical geometry and is effectively flat. This means that, if rendered in certain ways, panoramas can offer an appealing basis for video-conferencing, as they provide a full 360° view of an environment in a single image, but they are also a convenient context to temporally

and spatially relate videos within large collections.

The work presented in this thesis then demonstrates the suitability of videos in panoramic context to transmit spatio-temporal information describing a remote location to enable telecommunication and spatio-temporal browsing. To support our research, we conducted a series of user studies investigating the proposed representation. We developed two distinct videos+context applications to enable portable video-conferencing and spatio-temporal browsing of large video-collection respectively, and used the two platforms to conduct our investigation. Results of our studies show that videos in panoramic context can successfully convey spatial and temporal information describing remote places, which in turn enhance spatio-temporal thinking and present the remote environment in a way that users can intuitively understand and act upon. We showed that our representation can be adopted for teleconferencing scenarios, performing comparably to expensive panoramic video system and better than conventional webcam-style video chats. At the same time, we proved that our representation outperforms common video-browsing tools in spatio-temporal browsing tasks.

The structure of this thesis reflects the various stages of the research. The investigation started with a comprehensive review of fundamental works to the research topic (Chapter 2), narrowing down the focal area of research to the six most relevant topics including VMC and long-distance communication, video acquisition, transmission and rendering, video+focus applications, 3D reconstruction and depth fusion. Chapter 3 introduced the reader to BEAMING [Con10], the main project that motivated the research and under which a large part of the development was done. Chapter 4 further discussed the BEAMING idea by documenting the development of two instances of the platform, highlighting some of the methodological contributions of this thesis. Chapters 5–7 described the experimental works of this thesis through a series of user studies, and each chapter focused on a particular property of the videos in panoramic context representation, introducing the substantive and methodological contributions of this thesis. Chapter 5 investigated the suitability of a single video in panoramic context for collaborative telepresence scenario, documenting at the same time the development of *PanoInserts*, a portable teleconferencing system. Chapter 6 explored the effect of multiple videos in context on user spatial and temporal understanding of a remote scene, and described *Vidicontexts*, a spatio-temporal browsing interface. Finally, Chapter 7 studied the effect of display type on users interfacing with multiple videos in panoramic context, while Chapter 8 related back the findings of each user study to the thesis overarching goal and research questions. This closing chapter summarises the work presented in this thesis. First, methodological and substantive contributions are described. Finally, potential directions for future work are established.

9.1 Contributions

The overarching goal of the research was to investigate how videos in context may be employed to convey spatial and temporal information describing a remote location and the dynamics within, and how well this representation can replace more sophisticated solutions. Chapter 1 introduced three central research questions that are components of this goal:

1. *Can spatially localised video be used to increase the spatial information transmitted during video mediated communication, and does this improve quality of communication between users and their spatial thinking?*

2. *Can videos in panoramic context be used to convey spatial and temporal information describing a remote place and the dynamics within, and does this improve users' performance in tasks that require spatial and temporal thinking?*

3. *Measured by spatio-temporal thinking, is there an impact of display type on reasoning about events within videos in panoramic context?*

The first two questions are concerned with investigating the visual quality of video+context representation. They address the central premise of whether videos in panoramic context may be applied both in real-time to enhance the richness of VMC (Question 1), and offline to enhance spatio-temporal reasoning of people during video-collection browsing tasks (Question 2). The final question is secondary to the focus of the overall research, and addresses how the immersion level of a display can affect the perception of the video+panoramic context representation.

This thesis made both substantive and methodological contributions. The substantive contributions consist of empirical findings concerning the application of videos in panoramic context to both VMC and spatio-temporal browsing. The methodological contributions concern the development of solutions to acquire 3D models of large environments, stream and render depth information, acquire and render panoramic imagery and videos, and the development of two videos in panoramic context interfaces.

9.1.1 Methodological Contributions

Presented in the following section, the substantive contributions made by the experimental research concern the application of videos in panoramic context to transmit spatio-temporal information of remote places and facilitate VMC and video browsing applications. However, to facilitate the work from which these contributions are derived, development of two distinct platforms was required. This development work included a portable surround teleconferencing system, called *PanoInserts*, and an interface to allow spatio-temporal browsing of video-collections, named *Vidicontexts*. Both platforms are the result of collaborative development efforts, and are detailed in Chapters 5 and 6 respectively.

Chapter 3 presented the technical details of BEAMING, the ICVE system that was developed over the course of this research. BEAMING allows remote communication between remote sites, providing a collaborative mixed-reality environment that grants symmetrical social affordance and sensory cues to all connected users whether they are locals or visitors. Other ICVE systems, such as DIVE [AFH+97], MASSIVE [GB95] or Blue-C [GWN+03] also support this application. However, the unique feature of BEAMING is that the platform abandons the symmetry of access to a shared virtual environment in which collaboration happens, and rather focuses on recreating, virtually, a real environment and having remote participants visit that virtual model.

During the development of BEAMING, two platforms instances have been created and demonstrated: the BEAMING platform one (BP1) and BEAMING platform two (BP2). Both platforms are

the results of a collaborative effort. However, for both platforms, solutions related to the acquisition and transmission of the destination to the visitor can be considered as methodological contributions of this thesis. With respect to the BP1 (see Section 4.1), the candidate has been the main developer of solutions to support surrounding and 2.5D video acquisition, rendering and transmission. To this aim, he has developed solutions to interface with the cameras, render their video streams and transmit the data over the network. Concerning the BP2 (see Section 4.2), the candidate has developed rendering solutions to efficiently render the large point clouds generating from a RGBD mapper at the visitor site, which include dynamic frustum culling on GPU. Additionally, he has been the main developer of solutions to stream and calibrate a network of webcams, and he has contributed to solutions to calibrate the environment reconstruction, 3D static models and video streams together. Finally, additional methodological contributions include the experimental task designs, which may be used and adapted for future studies (Chapters 5–7).

9.1.2 Substantive Contributions

The substantive contributions of the experimental work, documented throughout Chapters 5–7, directly address the three central research questions posed at the beginning of the thesis. The first question asked whether spatially localised video could be used to increase the spatial information transmitted during VMC, and consequently, does this improve quality of communication and users' and spatial thinking. Work aiming to address this question is concerned with the theory of spatiality in mediated telecommunication, which is the degree to which a system supports fundamental properties such as movement, distance, containment, topology and a shared frame of reference such as a Cartesian coordinate system [BGR+98]. A central hypothesis of this research is that, by increasing capture, transmission, and display of spatial information about a remote location, VMC may be enriched, and medium will be more able to convey a sense of space which is more similar to the one perceivable in the real world. Findings from the experiment on object-focused placement documented in Chapter 5 form the main contributions to this topic. The study revealed that the video in panoramic context representation does convey a higher sense of space than conventional webcam-based system, obtaining comparable results to the more sophisticated fully-panoramic video based system. As one means to foster spatial awareness in VMC is to transmit a panoramic representation of a space to a remote viewer [FGR04, CRG+02], this result confirms that video in panoramic context applications support spatiality. Additionally, the study revealed that users can intuitively understand and act upon our prosed representation. Therefore, the first research question may be answered affirmatively, with a caveat stressing the fact that this finding cannot yet be extended to the more general case of multiple videos in panoramic context, as only a single video scenario was tested in the study.

The second question asked whether videos in panoramic context could be used to convey spatial and temporal information describing a remote place and the dynamics within it, and consequently, does this improve users' performance in tasks that require spatial and temporal thinking. Similarly to the preceding question, this question is related with the theory of remote spatial awareness. Additionally, work aiming to address this question is also concerned with the theory of focus+context systems,

which are interfaces showing a subset of information in full detail within a wider context of surrounding lower-density detail [CKB09, BGBS02]. A central hypothesis of this research is that, by automatically organising a video-collection with respect to time and space presenting this vast amount of information in its original context, users' spatio-temporal cognitive load may be eased. Findings from the experiment on object-focused localisation and tracking documented in Chapter 6 form the main contributions to this topic. The study revealed that our video-collection+context representation has significant improvements to accuracy and completion time in visual search tasks compared to existing video browsing systems. Insights from the study showed that providing panoramic contexts makes spatio-temporal tasks easier and faster, effectively resulting in a high level of spatiality conveyed by the visual representation, which in turn improves spatial awareness and eases users cognitive tasks' load. Hence, in accordance with our initial hypothesis, the second question may be answered affirmatively, extending the results obtained while investigating the first research question.

The third and final question asked whether, measured by spatio-temporal thinking, display type may be an impact factor while reasoning about events within videos in panoramic context. This question is related to a theory grounded in the domain of virtual environments. Virtual reality research has established that immersive displays, such as large FoV flat displays or HMDs, can improve user performance in tasks that require a high level of spatial reasoning or in tasks that mimic the real world [PCS+00, MJSS02, TGSP03]. The early work of Slater focuses on how immersive displays might afford users a greater sense of presence [SU93, SLU+96, SSA+01], and his studies discover that immersion can lead to increased performance in 3D spatial tasks. Therefore our initial hypothesis, in accordance with previous studies, was that the level of immersion of a display type could be a significant factor on users spatio-temporal thinking. Findings from the experiment on object-focused localisation and tracking documented in Chapter 7 form the main contributions to this topic. The study revealed that the level of immersion of a display, unlike the case of 3D environments exploration, cannot be considered a significant factor while reasoning about events within videos in panoramic context. While this result is in contrast with our initial hypothesis, the negative outcome can be actually be interpreted as a positive result for the videos in panoramic context representation. Finding from this experiment, in fact, allows us to extend the results from the studies presented in Chapters 5 and 6 to a variety of display types, including egocentric immersive HMDs. Additionally, the user study revealed interesting implications for designing panoramic imagery systems on different displays. For instance, we discovered that tablet displays, one of the display type considered during our study, were less effective than desktop displays even though participants felt just as capable. Hence, the third question, in contrast to previous questions, may be answered negatively, with a caveat stressing the importance that the relationship between the display type and its intended panoramic application should be carefully considered.

9.2 Limitations

The work presented in this thesis made both substantive and methodological contributions, as outlined in the previous section. While through the work presented here we were able to answer all the research questions introduced at the beginning of this thesis, we are aware that alternative routes could have been

taken during the development, and that the one documented here presents some limitations, which we will outline in the rest of this section.

9.2.1 Methodological Limitations

The methodological contributions of this thesis concern the development of solutions to acquire 3D models of large environments, stream and render depth information, acquire and render panoramic imagery and videos, and the development of two videos in panoramic context interfaces.

Regarding the two panoramic interfaces developed in this thesis, perhaps the biggest limitation is that the context employed in both instances is limited to 2.5D omnidirectional imagery. While we have shown that this has indeed beneficial effects on a variety or tasks, it also limits the type of videos, and hence applications, that can be used. In order to obtain a correct alignment of the videos to the panoramic content, the footage needs to be captured from roughly the same optical center of the panorama. This means that the recording camera is given limited motion in the remote environment. This is clearly a limitation of both systems, which we plan to overcome in future development by introducing three-dimensional contexts, as detailed in Section 9.3. Additionally, while panoramas are available for many locations in the world, and simple tools on smartphones make panorama capture easy, our approaches still requires a panorama as we register each video individually to it. With only sensor orientation data or marker-based alignment, videos could still be coarsely aligned within an empty context, though existing videos rarely have embedded orientation data. Future work could explore stitching videos to each other to build a context. Further, even with a panorama, our solutions will fail if large changes have occurred in the environment between the panorama and videos. For instance, many historical videos may only partially match the environment as building development is likely to have occurred. Similarly, existing panoramas of meeting rooms may become obsolete if furnitures are changed, or interiors refurbished. Here, we would have to rely on inter-video homography estimation for times in the video which do not match the panorama, anchored between times which do match. With no visual similarity at all, again we could only rely on captured orientation data.

Focussing on PanoInserts, we have already mentioned how the registration technique and colour balance algorithm used in the system present technical limitations. However, both problems were tackled and solved during the development of Vidicontexts. On the other hand, the way videos are streamed and rendered in the current version of PanoInserts may pose some challenges if a substantially large number of phones is used. Here, the limitation is posed by the hosting machine, as receiving, decoding and rendering a large numbers of videos in real-time is virtually infeasible with current CPU architectures. While modern architectures are pushing the boundaries of what can be achieved with CPU-based software (e.g., the Intel Quick Sync on Sandy Bridge or later CPUs contain hardware to decompress five or more videos at once [Shi11]), currently a possible way to mitigate this would be to delegate the rendering to a GPU architecture. Another viable option would be to give the user the chance to "expand" clusters of videos, so that only videos which are currently covering an area of interests would be decoded and rendered on screen.

Similar scaling problems arise with the current implementation of Vidicontexts. While our system

can successfully handle several tens of videos at once, the performance tend to degrade linearly with the number of videos to decode. However, and especially when using the egocentric first-person view mode, selective rendering could be employed to render only the videos that overlap with the current user's view-point. An initial version of this was implemented by the candidate for the BP2, even though that was aimed at large point-clouds rendering. Alternatively, an "expansion" metaphor with which users can inspect clusters of videos, ignoring the rest of the collections, could be implemented. Another aspect that could be improved in the current version of the system is the fact that videos need to be pre-registered before replay. Making the registration interactive would open the possibility to have live video-streams embedded with pre-recorded footage, allowing users to either perform collaborative tasks, or to compare environments over time.

In general, many errors can affect the quality of video alignment to the context for the Vidicontexts interface, including failures and artefacts in panorama stitching, incorrect or badly synchronized sensor data and camera metadata, large deviations from the proxy geometry assumption and large dynamic objects. The problem of temporally consistent video alignment is difficult even for state-of-the-art vision systems, and improving this is important future work. However, we posit that this improvement would cause a relatively small functional improvement in our interface, and instead we try to show that a useful and wanted system is still possible under these conditions. Further, while orientation sensor data can be bad, it does provide a full fallback for cases where visual alignment will have difficulty, and modern smartphones produce fittings from sensor data that are acceptable for many video-collection+context applications. Finally, our examples and experiments do not use real data from community video websites, and many challenges remain to provide context for these varied collections. Nevertheless, our work demonstrates the promise of videos+panoramic-context techniques in general, and produces a visual descriptions with immediate benefits over existing solutions for both telecommunications and video collection exploration software for limited subsets of videos.

9.2.2 Substantive Limitations

The substantive contributions of the experimental work, documented throughout Chapters 5–7, consist of empirical findings concerning the application of videos in panoramic context to both VMC and spatio-temporal browsing. One possible limitations of the experimental design used during our study is that the novelty factors of our systems was partially overlooked while collecting data. In the three experiments documented here we never considered a within subject design when comparing different systems. We did this as we were concerned about possible learning effects for a single subject. However, we could have designed our experiments using a repeated-measurement approach, hence mitigating the learning effect and investigating the effect of different interfaces on the same subject.

For the experiment presented in Chapters 5, we only focused on users performances. However, we could have expanded our investigation to analyses other aspects of VMC collaboration, such as spatial references in dialogues, by analysing users' dialogues and spatial deixis. There is evidence in literature that shows how grounding the interaction through referential statements and gestures made in relation to objects of common interest facilitate spatially-aware collaboration [Fil82]. To this end, analysing

wheather this is the case also for our representation would certainly corroborate our evidences.

In general, the tasks used during our studies were designed to be as representative as possible of typical real-world usage of the proposed systems. While this allows to generalise the results to other scenarios other than the one tested, it is also true that the results could be integrated with other tasks that closely resemble activities that can reach the "full potential" of our systems. For instance, in the PanoInserts case, we could have conducted a series of real meetings to evaluate the quality of the systems and integrate this with the existing results. In Section 5.5.4 we give an overview of how this could be done. Similarly, for the Vidicontexts case, we could have designed an additional tasks that closely resembles the most critical applications supported by our system, such as video-surveillance or virtual tourism. In Section 6.5.3 we give a list of viable tasks that could have been tested. However, please note that we do not intend to *replace* the existing tasks but rather we suggest that *additional* testing could have been done to corroborate our results. As such, we reserve this additional investigation in future work.

9.3 Directions for Future Work

The work presented in this thesis offered a solution to the general case of aligning video content to the real world to transmit spatio-temporal information of remote location. We already discussed how panoramic imagery offers a visual representation that stands in between pure 2D video and fully 3D geometry. One direction for future work, then, is to replace the 2.5D panorama context with 3D models of a remote environment. A work similar in spirit has been already explored by Neumann *et al.* [NYH+03] with their Augmented Virtual Environments (AVE) system. The system presented an initial solution to the challenging problem of aligning imagery to 3D models. While analysing the limitations of their systems, the authors noted that the proposed system was unable to properly display objects that are not part of the model. For example, lamp poles, cars, and trees are projected onto the buildings and roads, and they look warped and distorted from other viewpoints. This explains how extending our proposed representation to the 3D case is not trivial, opening opportunities for interesting future research. For instance, localising videos within large 3D models at interactive rates is still an open problem, with only few solutions available [SLK11, SLK12]. Similarly, segmenting foreground objects, a mandatory task to properly render objects located at different depth, remains an unsolved problem.

Once 3D models can be used as a context, one possible research path would be to replicate the experiments proposed in this thesis and a) investigate the benefits of videos in 3D context similarly to what done for the panoramic case and b) compare the panoramic and 3D contexts to establish the effective benefits of one over the other. One limitation of using panoramic imagery as context is that the user viewpoint is limited to a relatively small area surrounding the center of the panorama. With 3D models this limitation would potentially be removed, improving the quality of the experience and, possibly, spatial awareness.

However, with more sophisticated representation of the scene, rendering becomes a critical point to asses. Therefore, potential investigation could compare different ways of rendering the content and the context, for instance assessing video based or point based rendering or a combination of the two, to identify the impact of rendering quality on user performances. The challenges here would be twofold,

mostly related to the vast amount of data to render. First, rendering dozens of videos on top of large 3D model will pose an engineering problem, as this tasks is likely to have a high computational cost. Second, the variety and amount of visual information to render may result to be confusing or overwhelming for the users, and therefore smart ways of blending the various data stream will have to be identified. To achieve this, further investigation into how heterogeneous data are perceived and processed in 3D environments would be required.

Another interesting line of research may come from the application side. Being able to automatically align videos to 3D models would allow us to develop and test augmented reality applications. We could extend our *Vidicontexts* interface to work on portable devices and directly in real environments. We could use the 3D geometry as the reference onto which embed videos previously recorded, and the live video feed from the device's back-facing camera as the context. In other words, we could enable live exploration of real environments augmented with pre-recorded videos through portable devices. As the underlying reference is a 3D model, users would not be constrained into a single location, but rather they would be able to explore the whole space exploiting the mobility of the portable device.

Building on the display type effect user study presented in Chapter 7, one interest exploration could be done in the immersive display domain. By employing two type of contexts, 2.5D panoramic imagery and 3D models, and different display types, we could investigate the combined effect of context and display type with respect to easy of capturing, display affordability, system's presence and spatiality and user's performances.

Finally, interesting questions arisen from the discussion on the the results of our studies, and it would be valuable to investigate them in future work. First, in future development we would like to test our video-conferencing system to similar interfaces that exist in literature (e.g., the CamBlend system by Norris *et al.* [NSQ12]). Second, it would be interesting to extend the results obtained in Chapter 7 to include less conventional, but more immersive displays, such as large FoV projection displays or CAVE systems. Another interesting point of investigation could come from testing our interfaces for in-situ exploration of augmented environments. Finally, experimenting with the number of video insets used during remote teleconferencing could also reveal interesting aspects on the usefulness of (many) videos in panoramic contexts for telecommunications.

9.4 Conclusion

This thesis has aimed to investigate the use of videos in panoramic context to enhance teleconferencing and video browsing applications and improve user's spatio-temporal awareness. Research covered literature investigation, ICVE and video+context systems design and development, and different user studies covering both object-focused localisation and object-focused placement scenarios. The findings suggest that using videos in panoramic context allows to efficiently transmit spatio-temporal information describing a remote location, improving telecommunication and spatio-temporal browsing. Users interfacing with our proposed representation are able to achieve a high level of spatial awareness while performing remote spatial localisation tasks. Also, these findings are independent from the display type used, making the video+context representation suitable for a variety of displays and applications. Future

work will build on these findings by exploring the possibility to replace panoramic imagery with 3D models, assessing the benefits of doing so and exploring novel application scenarios.

Appendices

Appendix A

Publications

The following publications, all appearing in peer-reviewed international conferences and journals, are presented in chronological order according to date of publication. Where appropriate, the sections in this these corresponding to the work presented in the publication are referenced.

27^{th} Symposium on User Interface Software and Technology (UIST 2014)

Jie Song and Gábor Sörös and Fabrizio Pece and Sean Fanello and Shahram Izadi and Cem Keskin and Otmar Hilliges

In-air Gestures Around Unmodified Mobile Devices.

Conference on Visual Media Production, 2014

Fabrizio Pece, James Tompkin, Hanspeter Pfister, Jan Kautz and Christian Theobalt

Device Effect on Panoramic Video+Context Tasks.

Features extracts of work presented in Chapter 7.

26^{th} Symposium on User Interface Software and Technology (UIST 2013)

James Tompkin and Fabrizio Pece and Rajvi Shah and Shahram Izadi and Jan Kautz and Christian Theobalt

Video Collections in Panoramic Contexts.

DOI = 10.1145/2501988.2502013

Features extracts of work presented in Chapter 6.

SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)

Fabrizio Pece and William Steptoe and Fabian Wanner and Simon Julier and Tim Weyrich and Jan Kautz and Anthony Steed

PanoInserts: Practical Spatial Teleconferencing.

DOI = 10.1145/2470654.2466173

Features extracts of work presented in Chapter 5.

Presence: Teleoperators and Virtual Environments - 21(4), Fall 2012

William Steptoe and Jean-Marie Normand and Oyewole Oyekoya and Fabrizio Pece and Elias Giannopoulos and Franco Tecchia and Anthony Steed and Tim Weyrich and Jan Kautz and Mel Slater

Acting in Collaborative Multimodal Mixed Reality Environments.

Features extracts of work presented in Chapter 4

IEEE Computer Graphics and Applications, 2012

Anthony Steed and William Steptoe and Oyewole Oyekoya and Fabrizio Pece and Tim Weyrich and Jan Kautz and Doron Friedman and Angelika Peer and Massimiliano Solazzi and Franco Tecchia and Massimo Bergamasco and Mel Slater

Beaming: An Asymmetric Telepresence System.

DOI = 10.1109/MCG.2012.110

Features extracts of work presented in Chapter 3.

Theory and Practice of Computer Graphics - 2012

Fabian Wanner and Fabrizio Pece and Jan Kautz

Simplified User Interface for Architectural Reconstruction.

Conference on Visual Media Production, 2011

James Tompkin and Fabrizio Pece and Kartic Subr and Jan Kautz

Towards Moment Imagery: Automatic Cinemagraphs.

Joint Virtual Reality Conference of EGVE - EuroVR, 2011

Fabrizio Pece and Jan Kautz and Tim Weyrich

Adapting Standard Video Codecs for Depth Streaming.

DOI = 10.2312/EGVE/JVRC11/059-066

Features extracts of work presented in Chapter 4 and Appendix C.

Conference on Visual Media Production, 2010

Fabrizio Pece and Jan Kautz

Bitmap Movement Detection: HDR for Dynamic Scenes.

Additionally, during the doctoral study for this thesis, the candidate also contributed to additional juried exhibitions and workshops:

Discovery Zone. Luxembourg City Film Festival 2014

Jeff Desom, James Tompkin and Fabrizio Pece
Rear Window.

Symposium on User Interface Software and Technology (UIST 2014)
Demonstrations Session

Jie Song and Gábor Sörös and Fabrizio Pece and Sean Fanello and Shahram Izadi and Cem Keskin and Otmar Hilliges
In-air Gestures Around Unmodified Mobile Devices.

Symposium on User Interface Software and Technology (UIST 2013)
Demonstrations Session

James Tompkin and Fabrizio Pece and Rajvi Shah and Shahram Izadi and Jan Kautz and Christian Theobalt
Video Collections in Panoramic Contexts.

First Beaming Workshop, 2011

Fabrizio Pece and Jan Kautz and Tim Weyrich
Three Depth-Cameras Technologies Compared.
Features extracts of work presented in Chapter 3.

Finally, during the doctoral study for this thesis, the candidate has received the following awards and prizes:

- **Honorable Mention Award at CHI 2013** for PanoInserts paper;
- **Rabin Ezra Scholarship** 2010-2011 & 2011-2012;

and he has been invited to give the following talks:

- ETH Zürich Visual Computing Lunch - Zürich, Switzerland, Dec. 2013
- Dagstuhl Seminar on Real-World Visual Computing - Dagstuhl, Germany, Oct. 2013
- Max-Planck-Institut für Informatik - Saarbrcken, Germany, Oct. 2013
- BBC Research and Development - London, United Kingdom, June 2010

Appendix B

List of Acronyms

The following acronyms appear in this thesis:

2D	Two Dimensional
2.5D	Two-and-Half Dimensional
3D	Three Dimensional
3DTV	Three Dimensional Television
ANOVA	Analysis Of Variance
AR	Augmented Reality
BEAMING	Being in Augmented Multi-Modal Naturally Networked Gatherings
BP1	Beaming Platform One
BP2	Beaming Platform Two
BSS	BEAMING Scene Server
CAVE	CAVE Automatic Virtual Environment
CCD	Charge-Coupled Device
CCTV	Closed-circuit television
CG	Computer Graphics
CMOS	Complementary Metal Oxide Semiconductor
CPU	Central Processing Unit
COP	Centre of Projection
CV	Computer Vision
CVE	Collaborative Virtual Environment

DARPA Defense Advanced Research Projects Agency

DIVE Distributed Interactive Virtual Environment

DLP Digital Light Processing

DoF Degree of Freedom

DSLR Digital Single-Lens Reflex

DTAM Dense Tracking and Mapping

FAST Features from Accelerated Segment Test

FoV Field of View

fps Frames-per-Second

FVV Free Viewpoint Video

GPS Global Positioning System

GPU Graphics Processing Unit

GUI Graphical User Interface

HCI Human-Computer Interaction

HD High Definition ("Full HD" indicates 1920×1080 pixels)

HMD Head Mounted Display

HRTF Head-Related Transfer Function

IBR Image-Based Rendering

ICP Iterative Closest Point

ICVE Immersive Collaborative Virtual Environment

IMAX Image Maximum

IMU Inertial Measurement Unit

IR Infra-red

JPEG Joint Photographic Experts Group

JVT Joint Video Team

LIDAR Portmanteau of Light and Radar

ME Mean Error

MIT	Massachusetts Institute of Technology
MP	Megapixel
MPEG	Moving Picture Experts Group
MVS	Multi-View Stereo
MVV	Multi-View video
NCC	Normalised Cross Correlation
NITE	Natural Interaction Technology for End-user
NVC	Non-Verbal Communication
PBR	Point-Based Rendering
PC	Personal Computer
PERCRO	PERCeptual RObotics Laboratory
PSNR	Peak Signal-to-Noise Ratio
RANSAC	RANdom SAmple Consensus
RGB	Red Green Blue
RGBD	RGB-plus-Depth
SAD	Sum of Absolute Difference
SBI	Suppression of Background Illumination
SC	Sensory-Motor Contingency
SfM	Structure from Motion
FoV	Software Field of View
SIFT	Scale-Invariant Feature Transform
SL	Structured Light
SLAM	Simultaneous Localization And Mapping
SUS	Standard Usability Scale
SURF	Speeded Up Robust Features
TDMA	Time Division Multiple Access
ToF	Time of Flight

TUM	Technical University of Munich
UB	University of Barcelona
UCL	University College London
UI	User Interface
UDP	User Datagram Protocol
USB	Universal Serial Bus
VBO	Virtual Buffer Object
VCEG	Video Coding Experts Group
VDTM	View Dependent Texture Mapping
VE	Virtual Environment
VGA	Video Graphics Array
VMC	Video-Mediated Communication
VR	Virtual Reality
VRPN	Virtual Reality Peripheral Network

Appendix C

Streaming Depth

In this appendix we will report the results of our proposed depth-map compression algorithm obtained on a variety of depth-plus-colour videos acquired with a Microsoft Kinect unit. The results presented in this appendix complement the discussion introduced in Section 4.1.1. Please note that some of the images used in this chapter are adapted from the author's own work [PKW11].

C.1 Depth-map Compression Results

We tested three dynamic sequences with a number of frames between 300 and 450 (for each test all the frames have been used to compute the evaluation metrics), and with a resolution of 640×480 pixels. As quality metrics we decided to compute the Peak Signal-to-Noise Ratio (PSNR) and the absolute value of the mean error (ME). To integrate the results analysis we also show point-cloud renderings of the depth maps before and after the transmission.

For comparison purpose, we implemented two depth encoding schemes based on "bit multiplexing". In both cases we split the original 16-bit buffer in three chunks with varying sizes, but never bigger than 8 bits, and we then pack them in a three-channel image. In the first case (which we will call BIT1) we interleave the original bit sequence with the scheme shown in Figure C.1. For the second case (which we will call BIT2) we store the first six most important bits in the first six most important bits of the first channel, the subsequent five bits in the five most important bits of the second channel, and the final five bits in the five most important bits of the third channel. We then pad the remaining bits with zeros. In our tests we decided to employ both JPEG and VP8/H.264 compression to show the results of our encoding scheme with different compression techniques. While JPEG's compression is purely based on the image statistics, VP8 and H.264 encoders take advantage of both temporal and spatial properties of the input sequence.

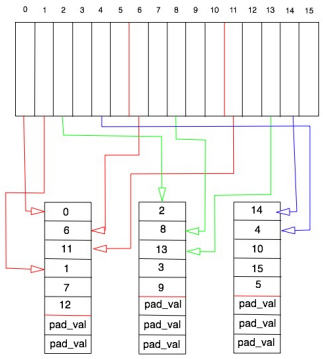
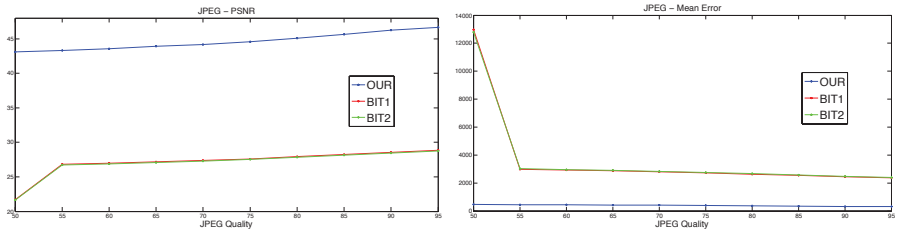


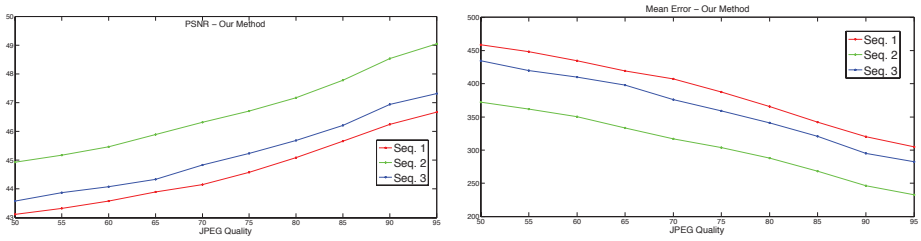
Figure C.1: *BIT1 interleaving scheme. Please note that each value in the 8-bit variable cells refers to the corresponding bit index in the 16-bit variable.*

C.1.1 JPEG Compression



(a) *PSNR* **(b)** *ME*

Figure C.2: *Results of the different depth encoding schemes using JPEG compression. Note how our encoding scheme yields a much better PSNR and a much lower ME. Results are computed on 450 frames with a resolution of 640×480 pixels.*



(a) *PSNR of our method* **(b)** *ME of our method*

Figure C.3: *Results of our technique using JPEG compression for the three sequences. 300–450 frames, 640×480 pixels.*

As first test, we combined our depth encoding scheme with the JPEG compression algorithm and compared our solution with the two bit-multiplexing schemes. Hence, we first encoded the video depth maps in an RGB image using either our compression algorithm or one of the other two schemes, then we applied JPEG compression with a certain quality level q, and finally we de-compressed the JPEG image

and decoded the resulting RGB into a single-channel, 16-bit map.

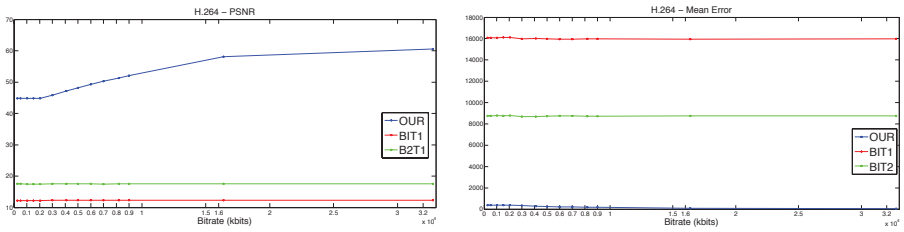
The result of this test, which we ran on the first video sequence, are shown in Figure C.2. The experiment has been conducted with increasing quality for the JPEG compression (quality level of 50 – 95). The performance of the proposed method is clearly superior to the bit-multiplexing schemes. Both PSNR and ME show how our method is able to compress and decompress the depth range without losing much precision. These results are also supported by the analysis of a point cloud of one of the compressed depth maps. Figure C.9 shows the decoded depth maps obtained with the three methods. The depth maps processed with our method are superior to the ones obtained with the bit-multiplexing schemes. In fact, while bit multiplexing leads to many grossly corrupted depth values, the quality of the depths obtained with our algorithm compares favourably to the ground truth. These results are confirmed by the tests run on the other two sequences (Figures C.3 and C.10, second column).

C.1.2 Video Codecs

After testing for JPEG compression, we run other tests on our depth encoder using two of the most common codecs used for real-time streaming, VP8 and H.264. For these tests, and for both codecs, we have used the codec implementations included in *ffmpeg* [FFm09]. Both VP8 and H.264 perform a colour-space transformation (RGB to YUV422) before starting the frame encoding, with higher precision in the Y channel. To ensure that the information contained in $L(d)$ is transferred as accurately as possible, we pack the encoded triples $L(d)$, $H_a(d)$ and $H_b(d)$ into Y, U, and V channel, respectively, and feed them directly to the *ffmpeg* encoder. Similarly for the bit-multiplexing techniques, we distribute values over Y, U and V according to their significance. We encoded the depth as the most significant 8 bits in the Y channel, and the remaining bits in the chroma channels.

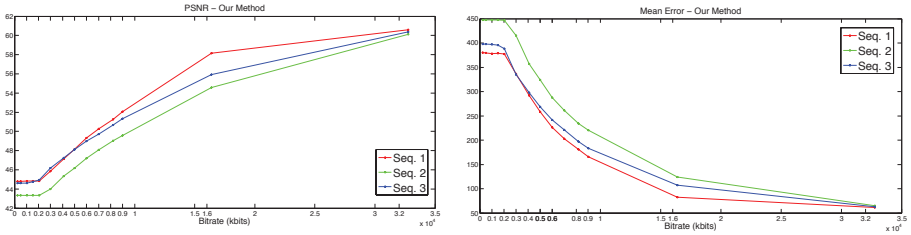
Note that all codecs considered (including JPEG) down-sample colour information spatially, which is another reason to store data of higher significance in the luminance channel. It further implies that our experiments also test for resilience to (moderate) spatial down-sampling and respective pre-convolution of the chromaticity of the image.

H.264 Codec



(a) *PSNR* **(b)** *ME*

Figure C.4: *Results of the different depth encoding schemes using H.264 compression. Note how our encoding scheme yields a much better PSNR and a much lower mean error. Results computed on 450 frames with a resolution of 640×480 pixels.*

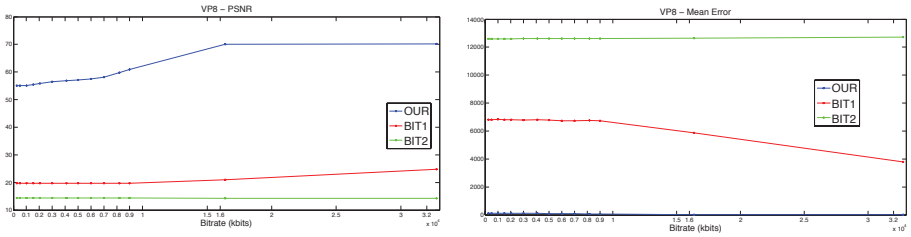


(a) *PSNR score of our method* **(b)** *ME*

Figure C.5: *Results of our technique using H.264 compression for the three sequences. 300–450 frames, 640×480 pixels.*

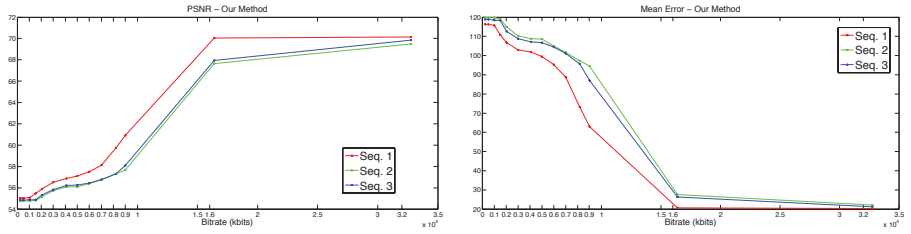
We started our video codecs experiment by combining our encoding scheme with the H.264 video compressor. Similarly to the JPEG case, the results of this experiment (Figures C.4 and C.5) revealed that our technique yields very good performance for both mean error and PSNR. Moreover, the amount of error introduced in the reconstructed maps do no seem to adversely affect the reconstructed depth maps (Figure C.10(h) and Figure C.10(l)). The overall scene's details are well preserved, and the error is mostly located around the edges.

VP8 Codec



(a) *PSNR* **(b)** *ME*

Figure C.6: *Results of the different depth encoding schemes using VP8 compression. Note how our encoding scheme yields a much better PSNR and a much lower mean error. Results computed on 450 frames with a resolution of 640×480 pixels.*



(a) *PSNR* **(b)** *ME*

Figure C.7: *Results of our technique using VP8 compression for the three sequences. 300–450 frames, 640×480 pixels.*

Finally, we run a test on the first of three sequences using our depth encoding scheme and the two bit-multiplexing techniques with VP8 compression. Figure C.6 shows the results of this initial test. The experiment has been conducted with increasing bit-rate (256 kbit–32768 kbit) using *ffmpeg* with default parameters. Our compression scheme yields the best performance for both PSNR and mean error, in contrast to the two bit-multiplexing techniques. Moreover, our method generates depth maps that are almost identical to the original ones (Figure C.10(c)). Figure C.7 shows the performance obtained by our algorithm for the other two video sequences, confirming the results of the previous test. The error introduced by our compression scheme is low, as is also clear from the point clouds showed in the third column of Figure C.10. From this, we can conclude that our solution can be used successfully with both VP8 and H.264 compression for depth streaming.

The results obtained during our tests show that the proposed solution successfully adapts standard video codecs to depth map streaming. Our solution requires negligible computational overhead and works well with several compression algorithms such as JPEG, VP8 and H.264. Limited amount of noise is introduced during compression, and the mean error shows that our method affects the depth values very little. The majority of the errors occupies the regions around depth discontinuities. This, however, has been already noticed in previous works [MMS+09, CSSH04, PHE+11], and thus it has to be expected when depth discontinuities are not dealt with separately.

(a) *Points filtering - View One*

(b) *Points filtering - View Two*

Figure C.8: *Initial decoded depth map (left) with outliers marked in red. Filtered point cloud of depth samples (right).*

These limitations can be partially solved by filtering the decoded depth maps, as shown in Figure C.8. Filtering these depth samples (left) based on local point-cloud density helps removing outliers and improves the quality of the reconstruction considerably (right).

(a) *Original – Sequence One* **(b)** *BIT1* **(c)** *BIT2* **(d)** *Our Method*

(e) *Original – Sequence Two* **(f)** *BIT1* **(g)** *BIT2* **(h)** *Our Method*

(i) *Original – Sequence Three* **(j)** *BIT1* **(k)** *BIT2* **(l)** *Our Method*

Figure C.9: *Comparison of reconstructed depth maps using different depth coding strategies and JPEG compression (75%).*

(a) *Original – Sequence One* **(b)** *JPEG – 75%* **(c)** *VP8 – 1024 kbits/sec* **(d)** *H.264 – 1024 kbits/sec*

(e) *Original – Sequence Two* **(f)** *JPEG – 75%* **(g)** *VP8 – 1024 kbits/sec* **(h)** *H.264 – 1024 kbits/sec*

(i) *Original – Sequence three* **(j)** *JPEG – 75%* **(k)** *VP8 – 1024 kbits/sec* **(l)** *H.264 – 1024 kbits/sec*

Figure C.10: *Depth maps reconstructed using our method. (Point cloud renderings.)*

Appendix D

"Videos in Context for Telecommunication" Experimental Material

D.1 Experiment Form and Questionnaires

This section includes material from the "Video in Context for Telecommunication" experiment outlined in Section 5. The following figures visualise the form and questionnaires filled by the participant at the end of the experiment.

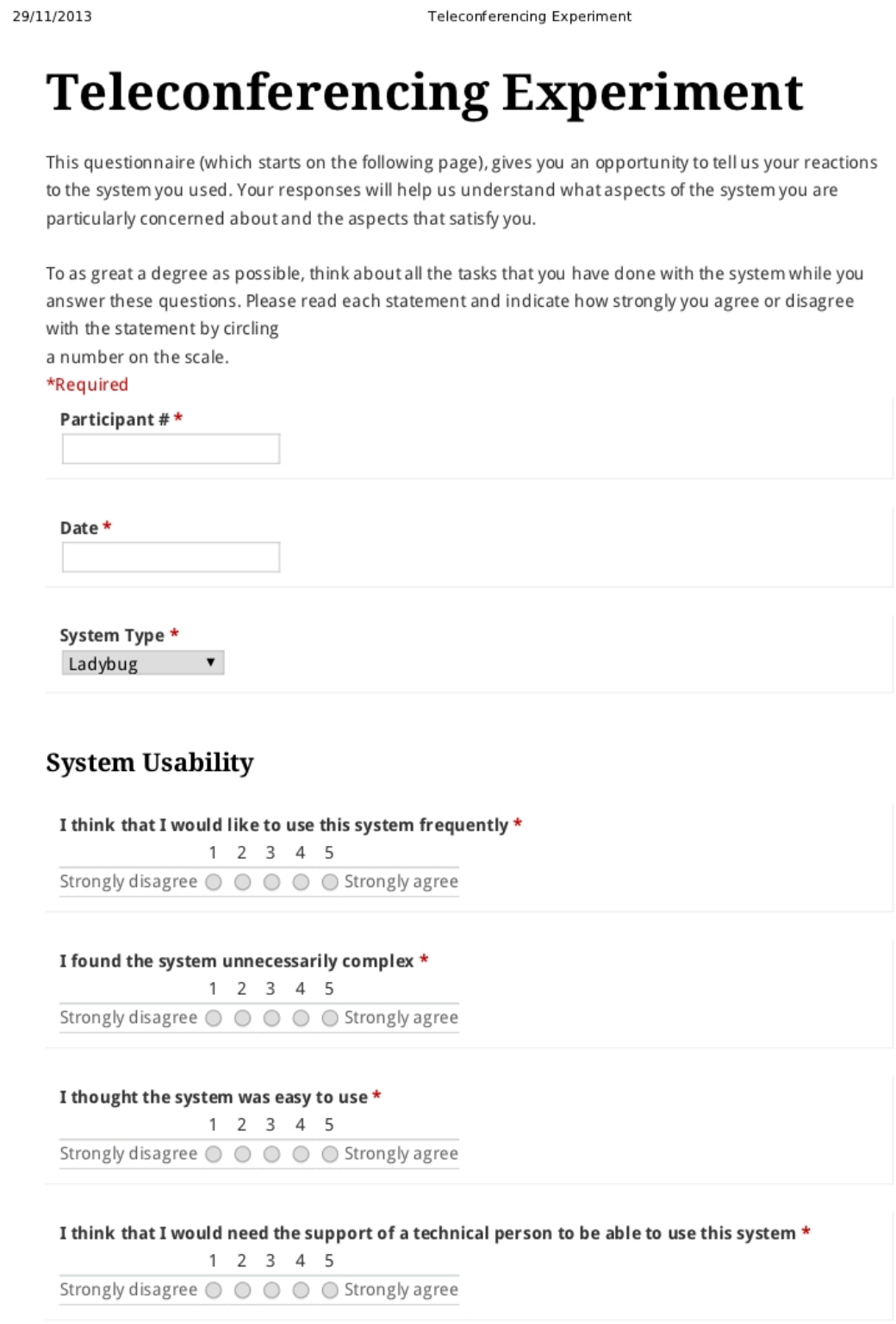


Figure D.1: *Experiment form.*

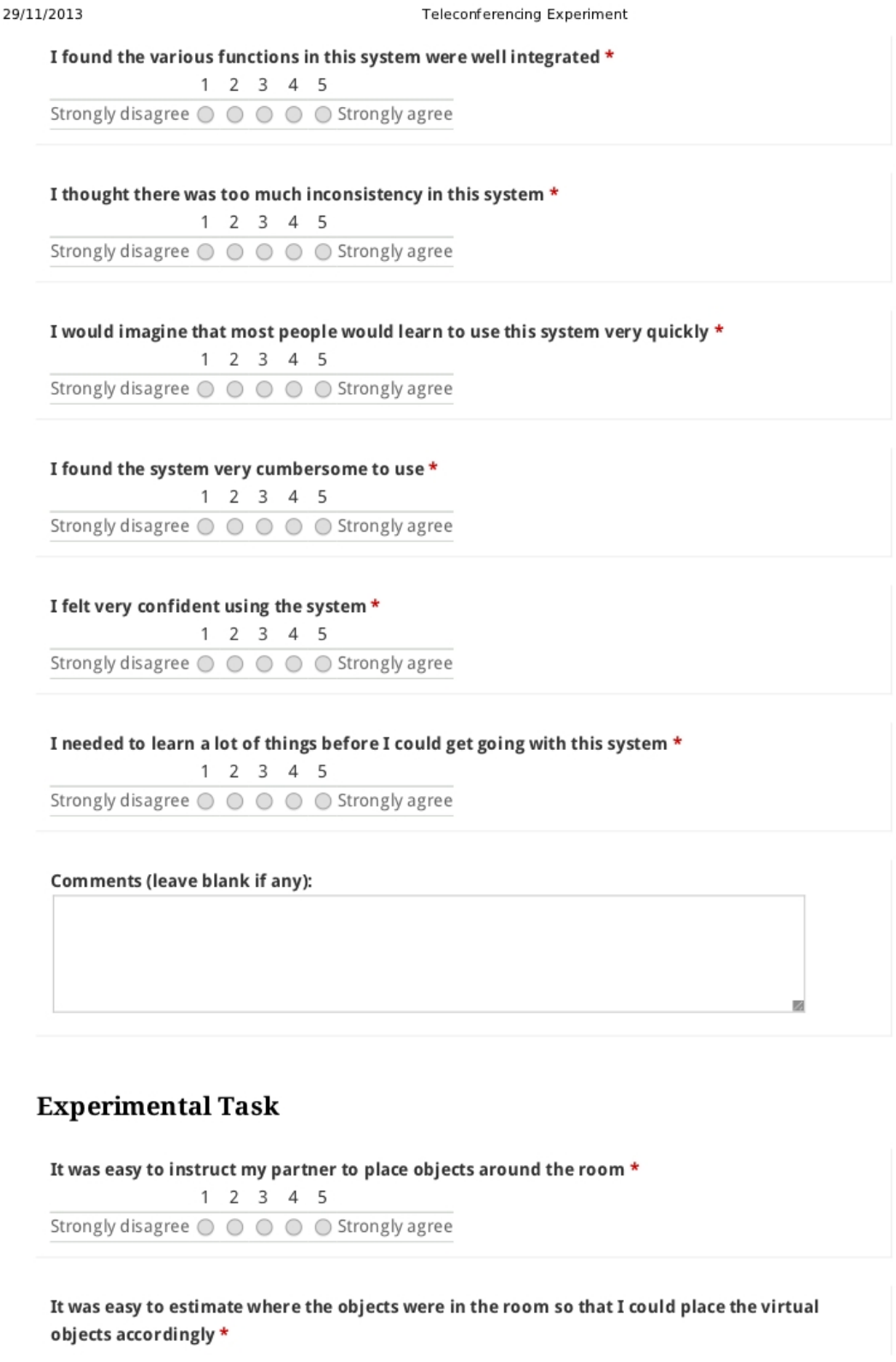


Figure D.2: *Experiment form (continued).*

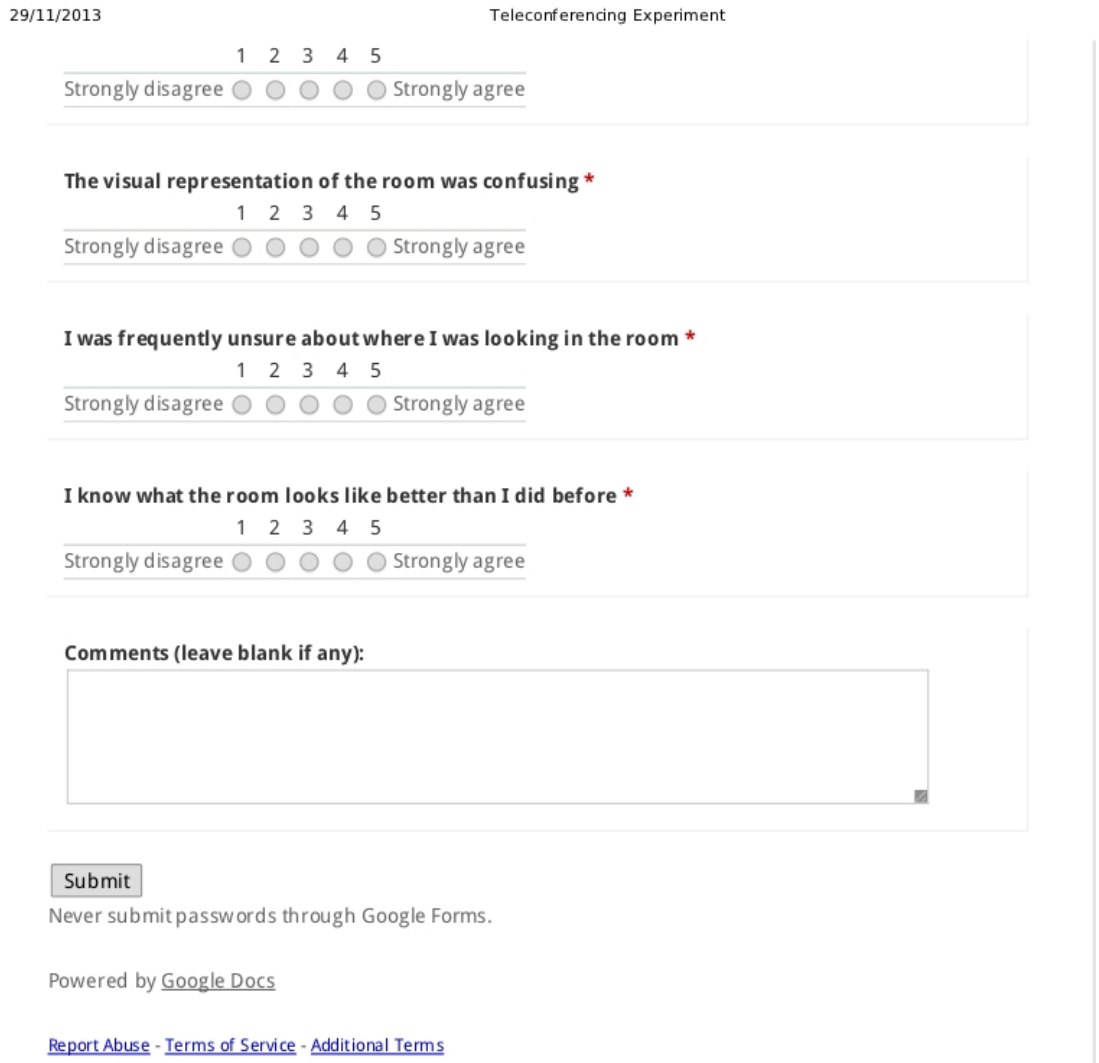


Figure D.3: *Experiment form (continued).*

Appendix E

"Videos in Context for Spatio-Temporal Browsing" Experimental Material

This section includes material from the "Videos in Context for Spatio-Temporal Browsing" experiment outlined in Section 6. Section E.1 visualises the baseline system used in our experiment, Apple iMovie. Section E.2 illustrates a MATLAB script that evaluates the validity of an homography, while Figures E.2–E.11 visualise the form and questionnaires filled by the participant during the experiment.

E.1 iMovie Interface

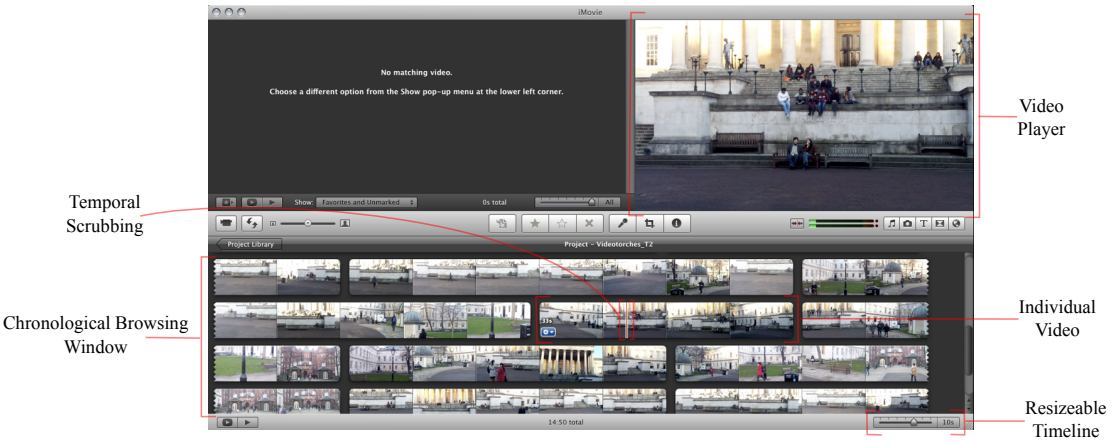


Figure E.1: *The iMovie interface used in our user study.*

To evaluate Vidicontexts, we decided to compare our system with *iMovie* as a baseline, and against iMovie with a panoramic context image available for reference (*iMovie+pano* henceforth). iMovie (Figure E.1) is consumer software typically used for non-linear video editing and, as its intended users are novices, it presents an intuitive interface. Its interface includes tools for browsing video collections and finding video content with which to edit. For our experiment, we ignore all of the editing features of iMovie and use only the intuitive video browsing tools. These tools are all accessible from the main window: 1) a chronological browsing area that displays videos as thumbnails and lets user skim through a video collection using hover scrubbing, 2) a resizeable timeline that can expand and contract the unit of time that each video thumbnail represents, and 3) a large panel used for video playback.

Initially, each video in the collection is presented as a single thumbnail and placed in chronological order in the browser. The user can expand the video into multiple thumbnails by using the timeline: coarser expansion values increase the time represented by each video thumbnail and so provides a collection overview, while finer values show more of the video as thumbnails and allow more time instances to be visible at once. Once a desirable video is found, the user can either select and play an entire video, or can hover the mouse over the thumbnails to scrub though that video section. iMovie offers additional functionalities for video editing, such as video cutting, which we did not use in our study.

In the iMovie+pano condition, users could also view a panoramic context for reference. The panorama of the scene was displayed at the same resolution as the one employed in Vidicontexts and in a separate window, and it was left to the user how they arranged their desktop space. All our users kept both iMovie and the panorama as full-screen windows and switched between having iMovie and the image viewer in the foreground. Most of our users switched back and forth throughout the tasks to view the reference image. Only a few users employed a different strategy: they viewed the context panorama once at the beginning of the task to obtain an idea of the surrounding space, and then focused only on the iMovie interface.

E.2 MATLAB Functions

```
function valid = validHomography(H)
% Test conditions for invalid homographies.
valid = true;
% Degenerate homographies.
N = 1000;
D = det(H);
if( D < 1/N || D > N )
    valid = false;
end
% Orientation reversing homographies.
A = H(1:2,1:2);
if( det(A) ≤ 0 )
    valid = false;
end
% Eigenvalue ratio is too large.
maxEigValRatio = 3;
[v w] = eig(A);
evRatio = max(diag(w))/min(diag(w));
if(evRatio > maxEigValRatio)
    valid = false;
end
% Foreshortening factor is too small;
% less than 1/3 along each direction.
if( w(2,2)*w(1,1) < (1/3).^2 )
    valid = false;
end
% Projectivity test.
maxVar = 0.01;
if( H(3,1).^2 + H(3,2).^2 < maxVar*maxVar)
    valid = false;
end
```

E.3 Experiment Form and Questionnaires

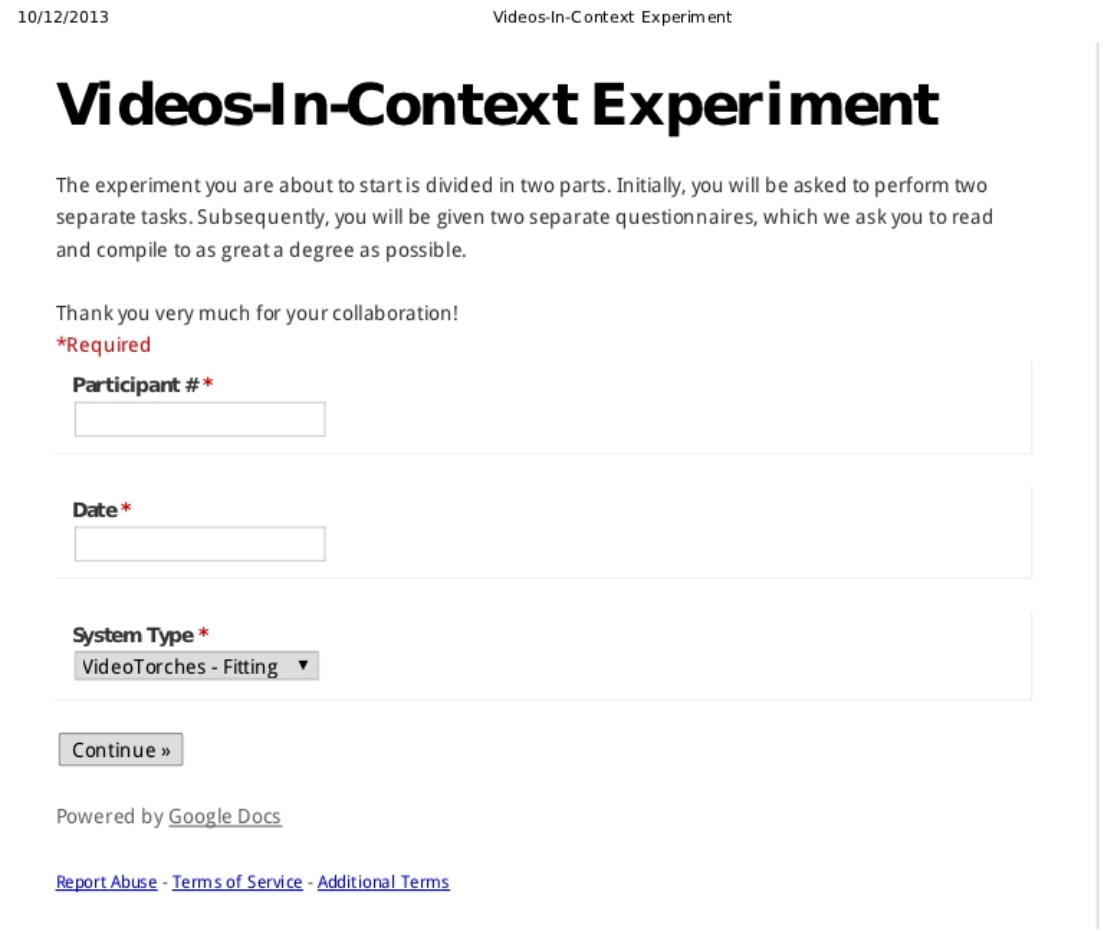


Figure E.2: *Experiment form initial page.*

10/12/2013 Videos-In-Context Experiment

Videos-In-Context Experiment

*Required

Task 1 - Counting

You will now be presented with a collection of videos which you can browse through the given interface. The videos have all been recorded in the same environment, a large outdoor square surrounded by architecture. Within this environment, there is a set of wooden benches located below the large building featuring a colonnade (see picture provided - Task1 - Counting: Picture).
Your task is to count how many DIFFERENT persons are sitting on these benches in the whole video collection. When done, please fill the input box below with your answer.
There is no time limit to this task, but please DO press the button "Start Task 1" below before to start the task, and then press again the same button when finished.

I have counted this many DIFFERENT people sitting on the benches *
Please provide numbers only

How familiar are you with the environment depicted (task one)? *
Familiar

« Back Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

Figure E.3: *Counting task briefing.*

Figure E.4: *Counting task image provided.*

10/12/2013 Videos-In-Context Experiment

Videos-In-Context Experiment

*Required

Task 2 - Tracking

You will now be presented with a second collection of videos, which you can browse through the given interface.
The videos have all been recorded in the same environment, a large outdoor square surrounded by architecture. Within this environment, there is a region bounded by a set of large, red umbrellas and a yellow door (see picture below: Task2 - Tracking: Picture).
Your task is to track different people in this area. Please count how many DIFFERENT persons can be seen entering, walking through and exiting the area of interest -- either from left to right or right to left. People exiting the area through the door can be counted in.
When done, please fill the input box below with your answer.
There is no time limit to this task, but please DO press the button "Start Task 2" below before to start the task, and then press again the same button when finished.

I have counted this many DIFFERENT people crossing the area of interest *
Please provide numbers only

How familiar are you with the environment depicted (Task two)? *
Familiar

« Back Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

Figure E.5: *Tracking task briefing.*

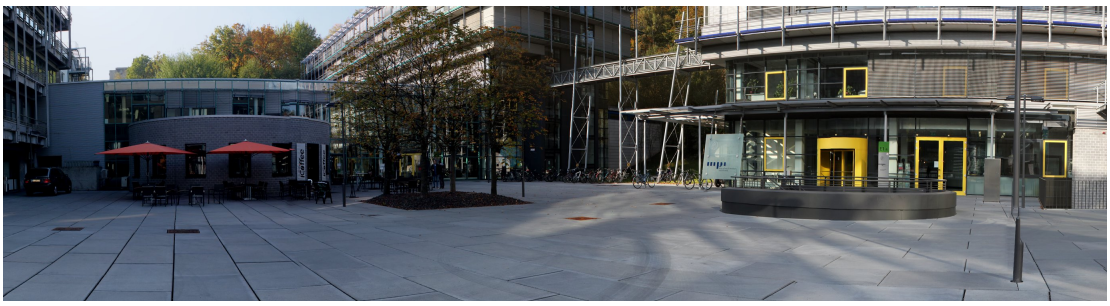


Figure E.6: *Tracking task image provided.*

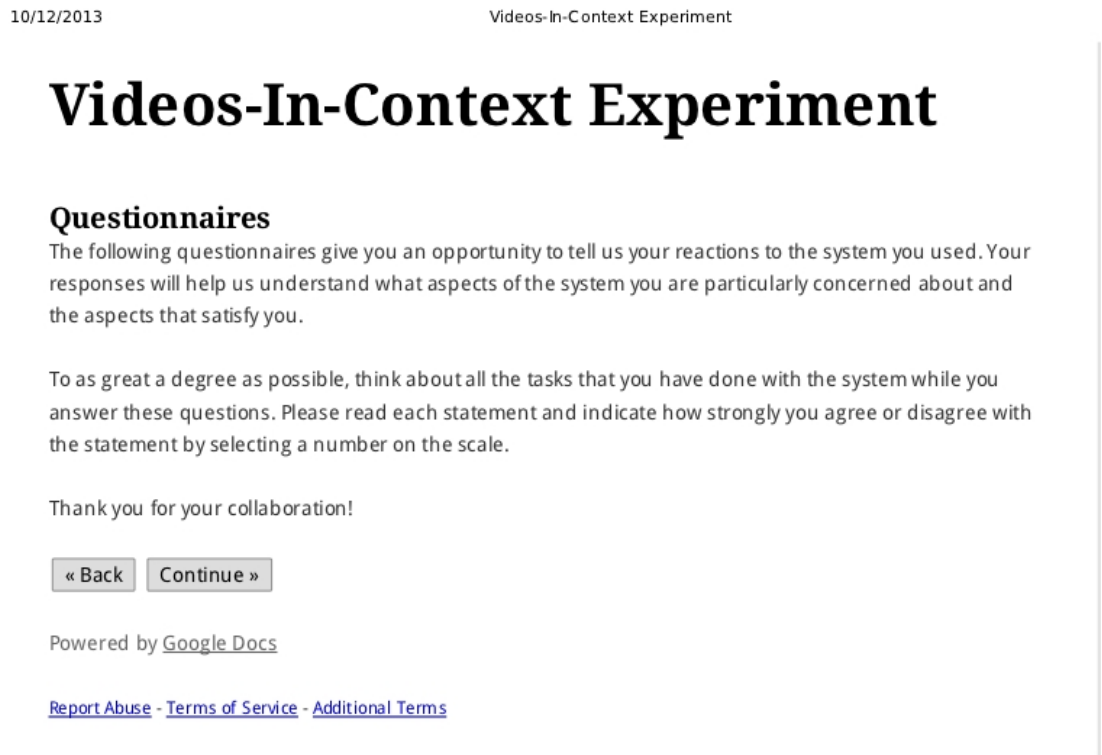
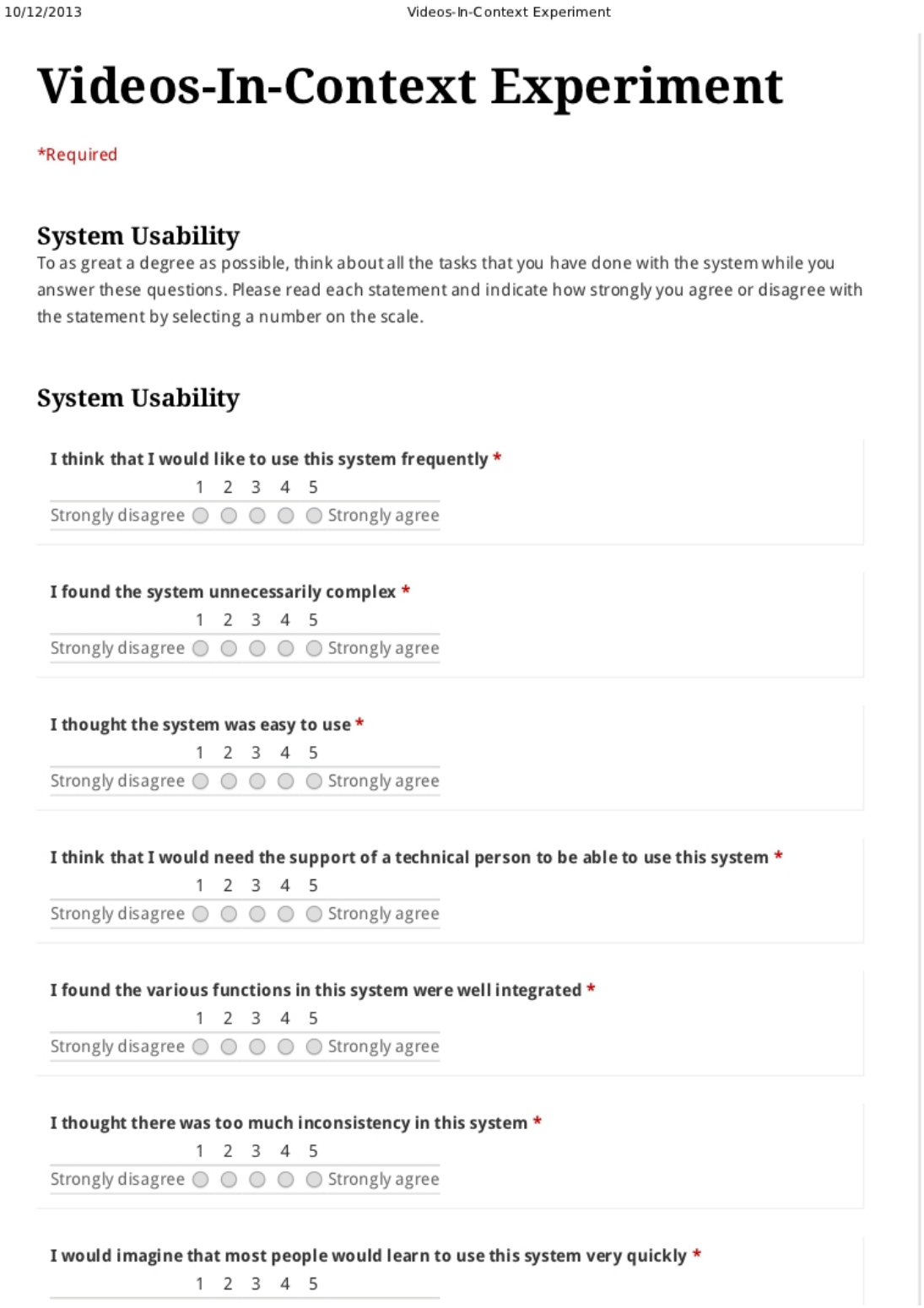


Figure E.7: *Questionnaires briefing.*



10/12/2013 Videos-In-Context Experiment

Videos-In-Context Experiment

*Required

System Usability

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions. Please read each statement and indicate how strongly you agree or disagree with the statement by selecting a number on the scale.

System Usability

I think that I would like to use this system frequently *

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I found the system unnecessarily complex *

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I thought the system was easy to use *

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I think that I would need the support of a technical person to be able to use this system *

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I found the various functions in this system were well integrated *

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I thought there was too much inconsistency in this system *

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I would imagine that most people would learn to use this system very quickly *

1 2 3 4 5

Figure E.8: *SUS questionnaire.*

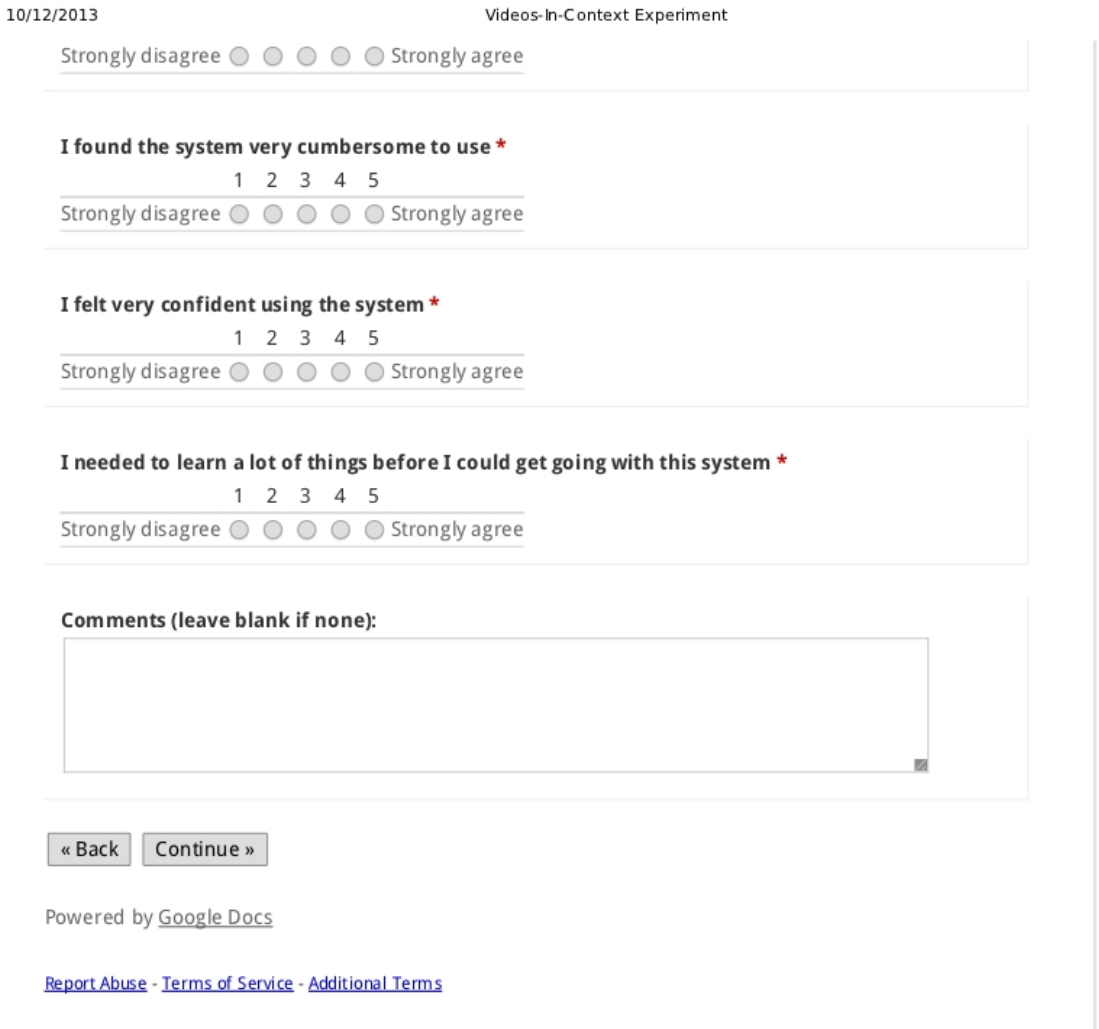
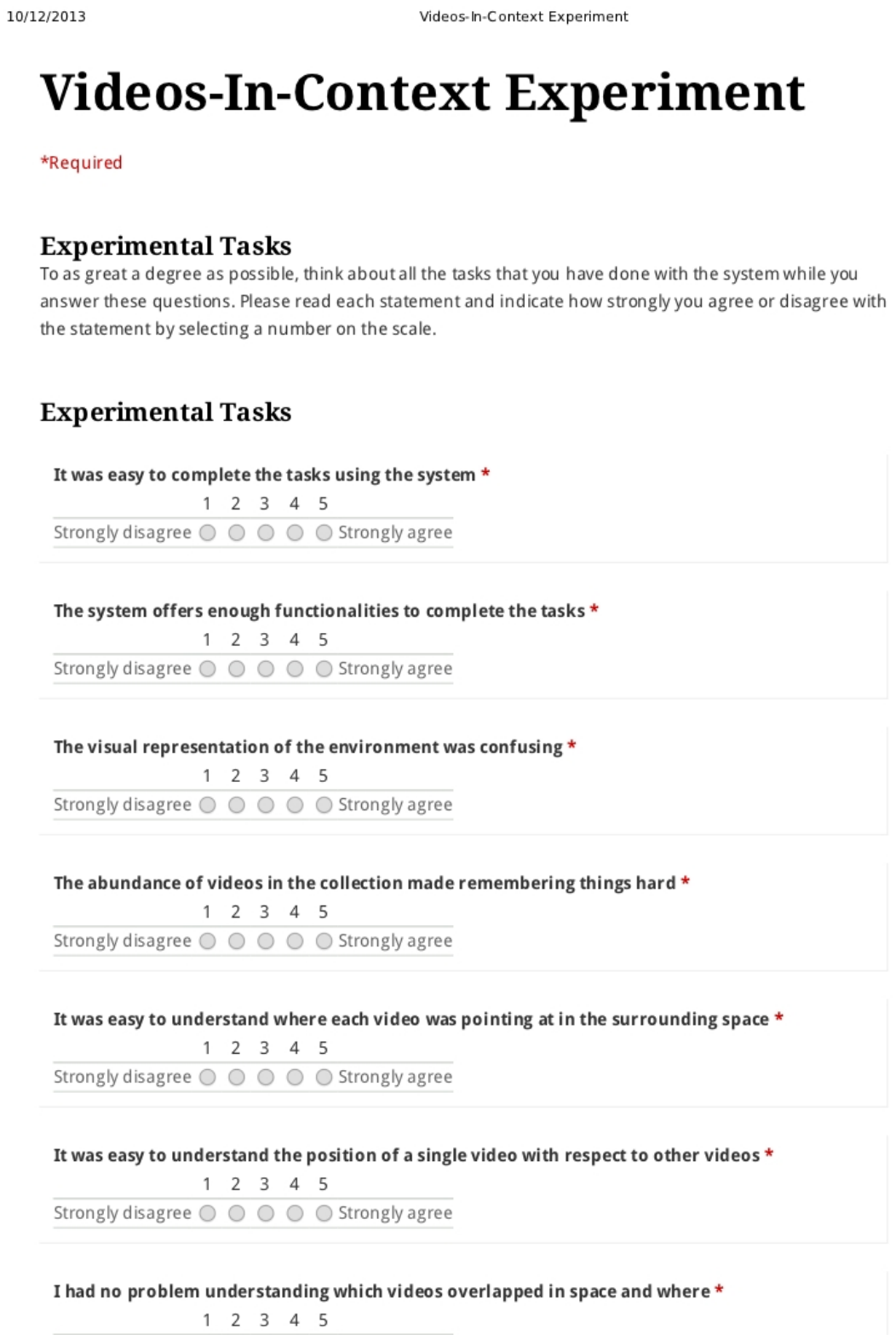


Figure E.9: *SUS questionnaire (continued).*



10/12/2013 Videos-In-Context Experiment

Videos-In-Context Experiment

*Required

Experimental Tasks

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions. Please read each statement and indicate how strongly you agree or disagree with the statement by selecting a number on the scale.

Experimental Tasks

It was easy to complete the tasks using the system *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The system offers enough functionalities to complete the tasks *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The visual representation of the environment was confusing *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The abundance of videos in the collection made remembering things hard *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

It was easy to understand where each video was pointing at in the surrounding space *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

It was easy to understand the position of a single video with respect to other videos *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I had no problem understanding which videos overlapped in space and where *

	1	2	3	4	5

Figure E.10: *Task-related questionnaire.*

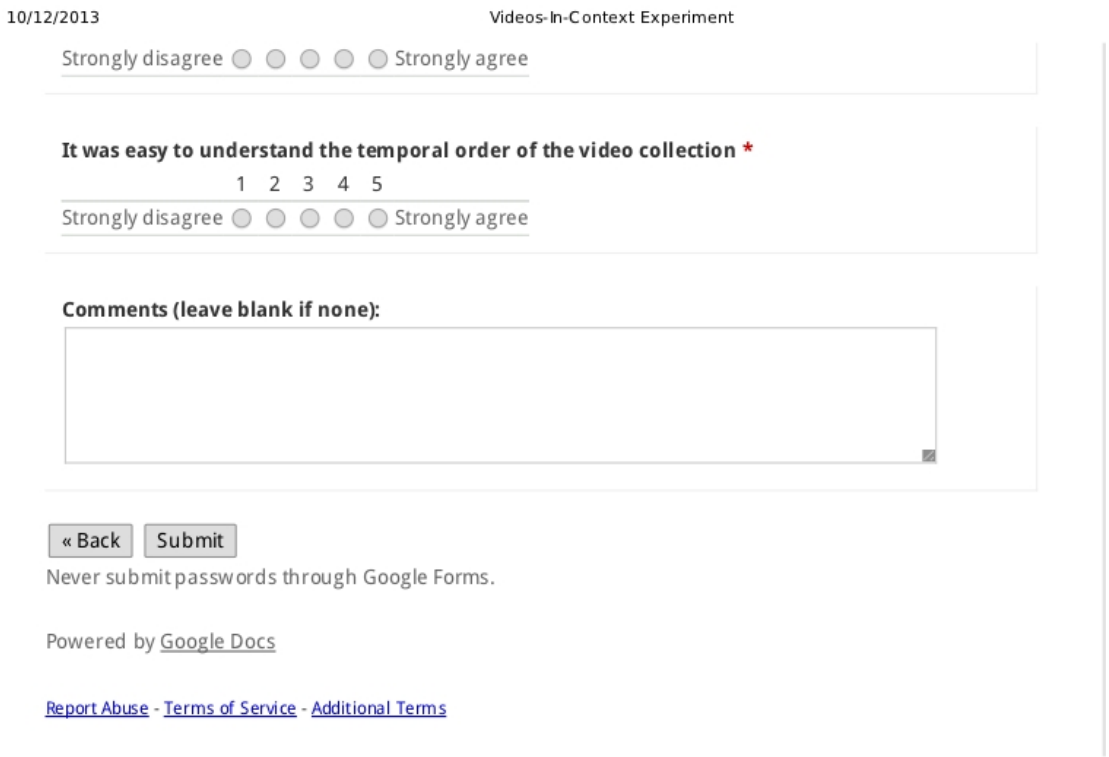


Figure E.11: *Task-related questionnaire (continued).*

Appendix F

"Immersive Display Effect on Videos in Panoramic Context Tasks" Experimental Material

This section includes material from the "Immersive Display Effect on Panoramic Videos in Context Tasks" experiment outlined in Section 7. Section F.1 presents and extension of our Vidicontexts interface to spherical display and augmented reality applications. Figures F.3–F.12 visualise the form and questionnaires filled by the participant during the experiment.

F.1 Additional Display Applications

The video-collection+context representation naturally fits display and interaction devices beyond desktop environments. We extend Vidicontexts to work on portable devices, such as tablets, HMDs and spherical displays. While our desktop interface shows either a perspective projection or an equirectangular projection, this exploration of display applications maps the panorama to both virtual and real spatially-located spheres. As details on the tablet and HMD extensions are reported in Chapter 7, this section will focus on spherical display and augmented reality applications only.

F.1.1 Spherical Interface

Figure F.1: *Additional displays and interactions.* Left: *A tablet acting as a proxy controller, where the spherical display mirrors the context of the tablet.* Centre: *Spherical display with a joypad controlling a cursor.* Right: *Tablet display in situ, showing a protest that no longer exist in the real environment.*

In this example, our context is displayed on a physical sphere, the Global Imagination's Magic Planet spherical display [Glo06], in tandem with complementary controller interfaces. Multiple users are able to

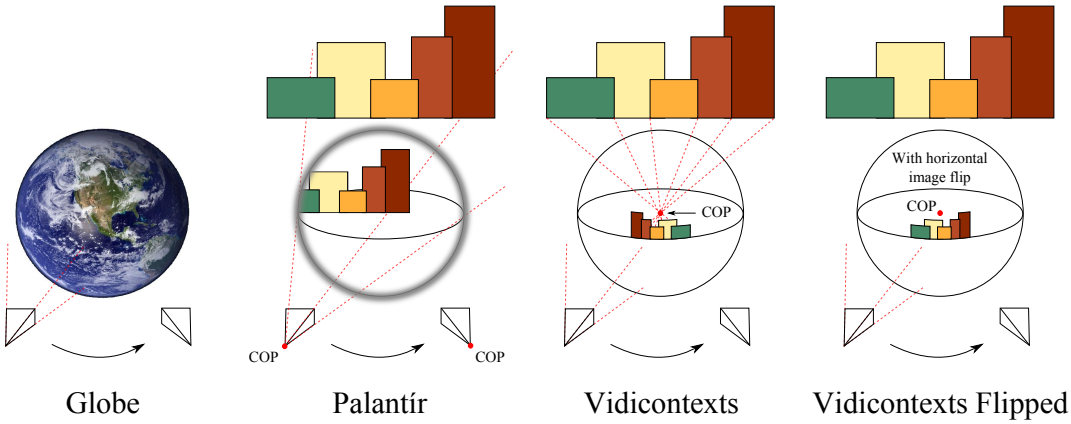


Figure F.2: Left: *World globe - if the user changes their viewpoint, then they will reveal content located on the far side of the sphere.* Center-Left: *Palantìr - changing viewpoint reveals different areas of the projected space similarly to what happens when moving past a window.* Center-Right *Vidicontexts - the projected world appears flipped left to right, and when moving to the right, the world to the left is revealed.* Right: *Vidicontexts with Flip: if we horizontally flip the image, when walking to the right, the world to the right is revealed.*

walk around the display to inspect different areas of the context and physically track videos as they move. Users can also control the system through a joypad or a tablet device, though a touch interface is also possible [BWB08]. With the joypad, users control a cursor on the spherical display, with modifications to exploit the specific controls, e.g., manipulating time in videos using the analog triggers (Figure F.1, *centre*). With the tablet complement, our existing flat display interface acts as a proxy controller, and any view changes on the tablet are reflected on the sphere (Figure F.1, *left*).

While mobile devices and HMDs naturally respect the geometry of inside→out video collections, the mapping to a spherical display requires more thought. Users observing a spherical display typically expect it to behave either as a) a world in miniature, such as a globe, or b) a magical seeing stone, or *palantìr*[1], which acts as a portal to another place or world. However, the Vidicontexts case is neither of these, as we explain here and in Figure F.2:

Globe: Content on the globe is mapped directly to the spherical display. If the user changes their viewpoint by walking around the spherical display, then they will reveal content located on the far side of the sphere (Figure F.2, *left*). Moving to the right reveals content on the globe farther to the right — the motion/content is consistent with the world in miniature.

Palantír: The sphere is a portal to a different place. Changing viewpoint reveals different areas of the space through the portal via parallax, similar to what happens when moving past a window (Figure F.2, *center-left*). The sphere boundary as seen from the viewer separates the two places, and the world is projected "through" the sphere to the eyes of the user. Thus, simulating a palantír with a spherical display and correctly rendering the panoramic context requires knowledge of the user's eye position. This could be discovered with an external head-tracking system, and this would limit the display to a single user.

[1] From *The Lord of the Rings* literary saga, by J. R. R. Tolkien.

Vidicontexts: The world to be viewed is projected onto the surface of a sphere, with center of projection at the center of the sphere. This is the creation of the panoramic context by photography. When the context is viewed with a tablet or HMD, the viewer is effectively in the center of the sphere. However, when we map this to the surface of a spherical display, we are now observing the world from *outside* – we have turned the world in on itself. There are two options for this projection:

1. **Flipped:** (Figure F.2, *center-right*) The world is projected onto the sphere. To maintain viewing directions, the world is projected onto the back of the sphere, that is, the sphere-ray intersection points which are farthest from the world when projected through the center of the sphere. When this projection is viewed from outside the sphere, the world appears flipped left to right. As the user walks around the sphere, the world is revealed as per the palantír case, where movement to the right reveals the world to the left. However, the whole world is horizontally flipped.

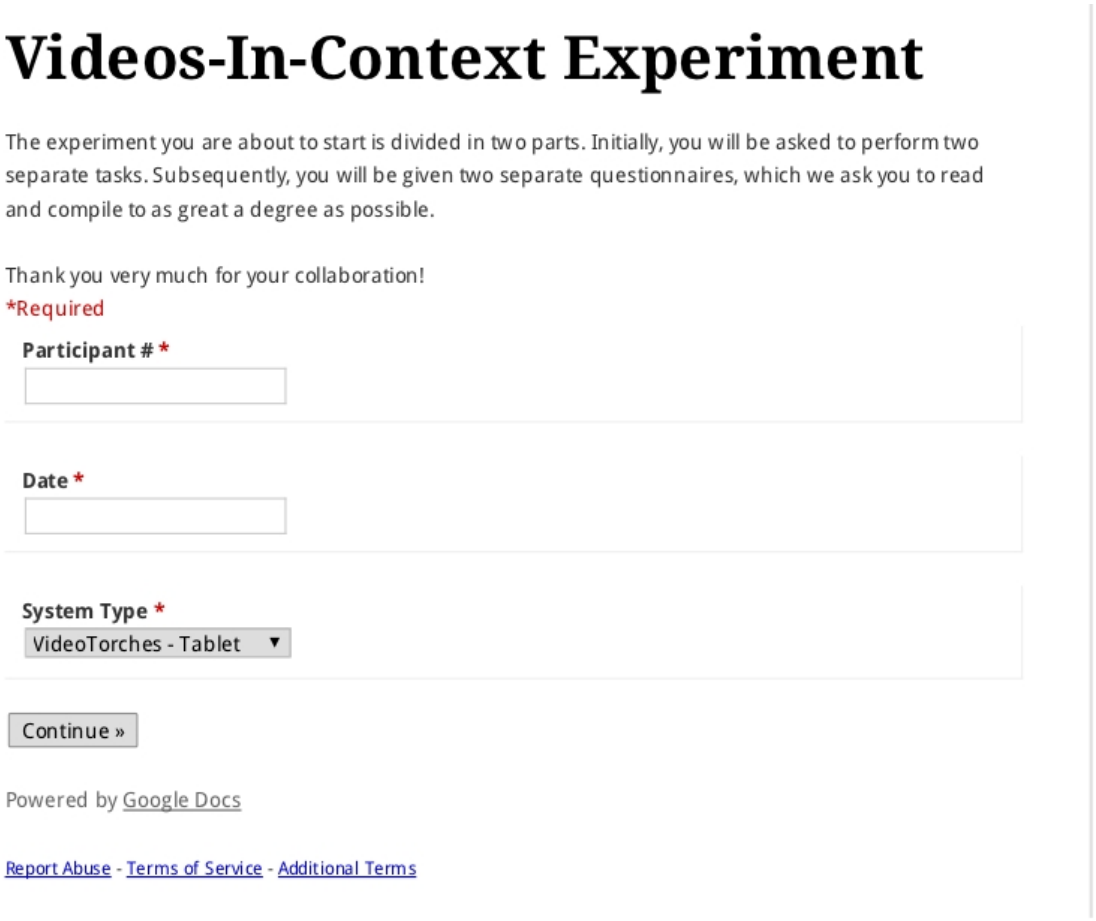
2. **No flip/bad parallax:** (Figure F.2, *right*) If we horizontally flip the image to try to correct this problem, the world appears correct from a single viewpoint. However, now, when walking to the right, the world to the right is revealed rather than the expected parallax effect of the world to the left being revealed.

Without head tracking, it is impossible to reconcile these two problems as we are viewing the world inverted. Either the world is horizontally flipped and movement around the spherical display is correct, or the world is not flipped and movement is inverted. The influence on users of these effects is not straightforward to understand or quantify. Future work should experimentally investigate the three options presented to try to estimate the impact on users perception and performance of these projection methods for spherically displaying inside→out video collections.

F.1.2 Augmented Reality

Our representation is also useful in augmented reality applications where the goal is to compare videos in situ using the real world as a context. This situation might occur as a curated experience at a cultural heritage site, or as a virtual tourism application where participants are GPS guided around a city and stand in hotspots to compare videos of past events with the current situation. GPS and orientation data are often sufficient for rough registration with the environment and, with this, in our example the user sees a protest in video that no longer exists in the real environment (Figure F.1, *right*). If a vision-based registration between mobile device and environment is required, with our representation the back-facing camera image need only be registered with the panorama once in real-time for all videos in the collection to be registered. In this case, the camera image would replace the panorama in our interface, though we leave this fine registration for future work.

F.2 Experiment Form and Questionnaires



Videos-In-Context Experiment

The experiment you are about to start is divided in two parts. Initially, you will be asked to perform two separate tasks. Subsequently, you will be given two separate questionnaires, which we ask you to read and compile to as great a degree as possible.

Thank you very much for your collaboration!

*Required

Participant # *

Date *

System Type *

VideoTorches - Tablet

Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

Figure F.3: *Experiment form initial page.*

Videos-In-Context Experiment

*Required

Task 1 - Counting

You will now be presented with a collection of videos which you can browse through the given interface. The videos have all been recorded in the same environment, a large outdoor square surrounded by architecture. Within this environment, there is a set of wooden benches located below the large building featuring a colonnade (see picture provided - Task1 - Counting: Picture).
Your task is to count how many DIFFERENT persons are sitting on these benches in the whole video collection. When done, please fill the input box below with your answer.
There is no time limit to this task, but please DO press the button "Start Task 1" below before to start the task, and then press again the same button when finished.

I have counted this many DIFFERENT people sitting on the benches *

Please provide numbers only

How familiar are you with the environment depicted (task one)? *

Familiar

« Back Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

Figure F.4: *Counting task briefing.*

Figure F.5: *Counting task image provided.*

Videos-In-Context Experiment

*Required

Task 2 - Tracking

You will now be presented with a second collection of videos, which you can browse through the given interface.
The videos have all been recorded in the same environment, a large outdoor square surrounded by architecture. Within this environment, there is a region bounded by a set of large, red umbrellas and a yellow door (see picture below: Task2 - Tracking: Picture).
Your task is to track different people in this area. Please count how many DIFFERENT persons can be seen entering, walking through and exiting the area of interest -- either from left to right or right to left. People exiting the area through the door can be counted in.
When done, please fill the input box below with your answer.
There is no time limit to this task, but please DO press the button "Start Task 2" below before to start the task, and then press again the same button when finished.

I have counted this many DIFFERENT people crossing the area of interest *

Please provide numbers only

How familiar are you with the environment depicted (Task two)? *

Familiar

« Back Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

Figure F.6: *Tracking task briefing.*

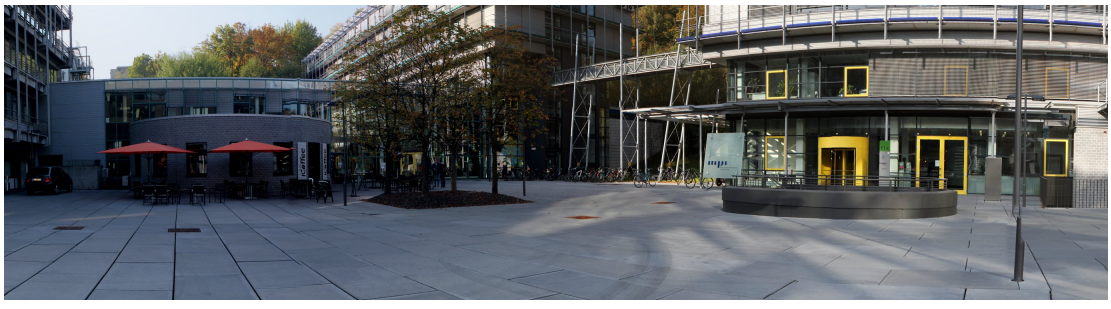


Figure F.7: *Tracking task image provided.*

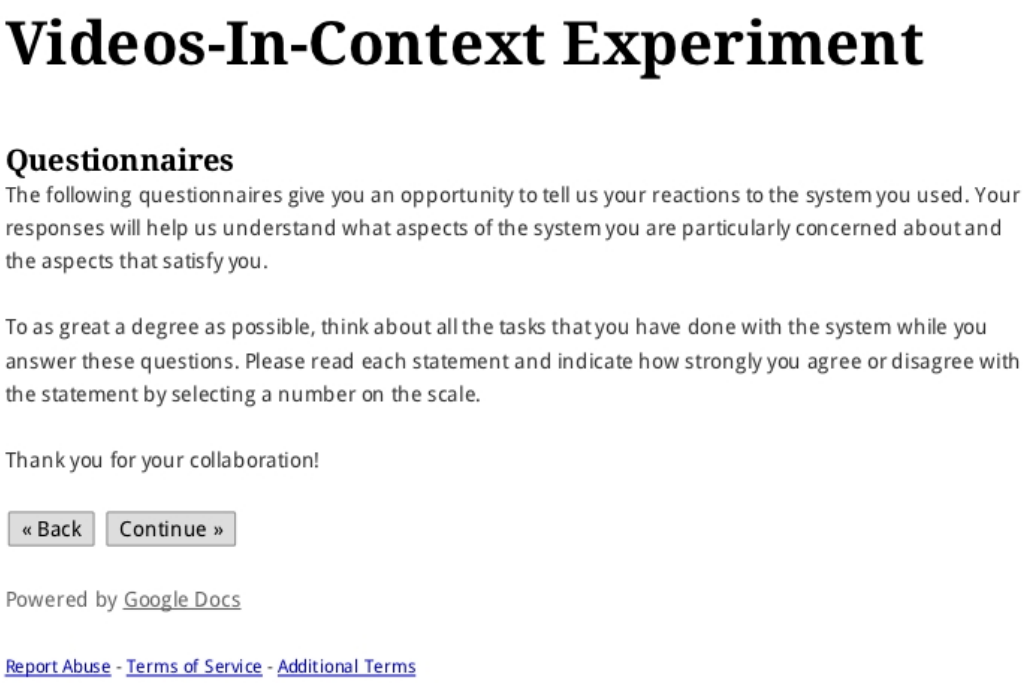


Figure F.8: *Questionnaires briefing.*

Videos-In-Context Experiment

*Required

System Usability

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions. Please read each statement and indicate how strongly you agree or disagree with the statement by selecting a number on the scale.

System Usability

I think that I would like to use this system frequently *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I found the system unnecessarily complex *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I thought the system was easy to use *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I think that I would need the support of a technical person to be able to use this system *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I found the various functions in this system were well integrated *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I thought there was too much inconsistency in this system *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I would imagine that most people would learn to use this system very quickly *

	1	2	3	4	5

Figure F.9: *SUS questionnaire.*

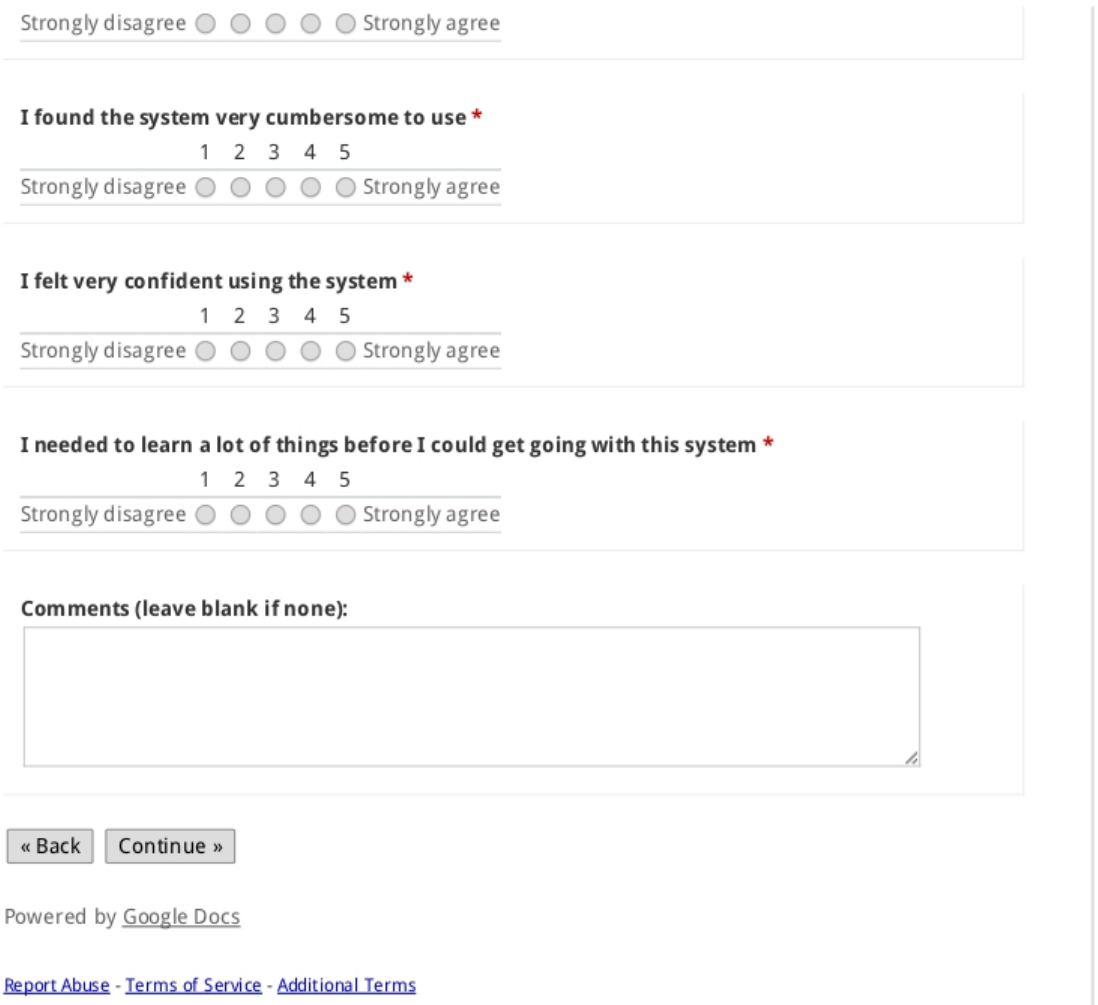


Figure F.10: *SUS questionnaire (continued).*

Videos-In-Context Experiment

*Required

Experimental Tasks

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions. Please read each statement and indicate how strongly you agree or disagree with the statement by selecting a number on the scale.

Experimental Tasks

It was easy to complete the tasks using the system *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The system offers enough functionalities to complete the tasks *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The visual representation of the environment was confusing *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

The abundance of videos in the collection made remembering things hard *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

It was easy to understand where each video was pointing at in the surrounding space *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

It was easy to understand the position of a single video with respect to other videos *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I had no problem understanding which videos overlapped in space and where *

	1	2	3	4	5

Figure F.11: *Task-related questionnaire.*

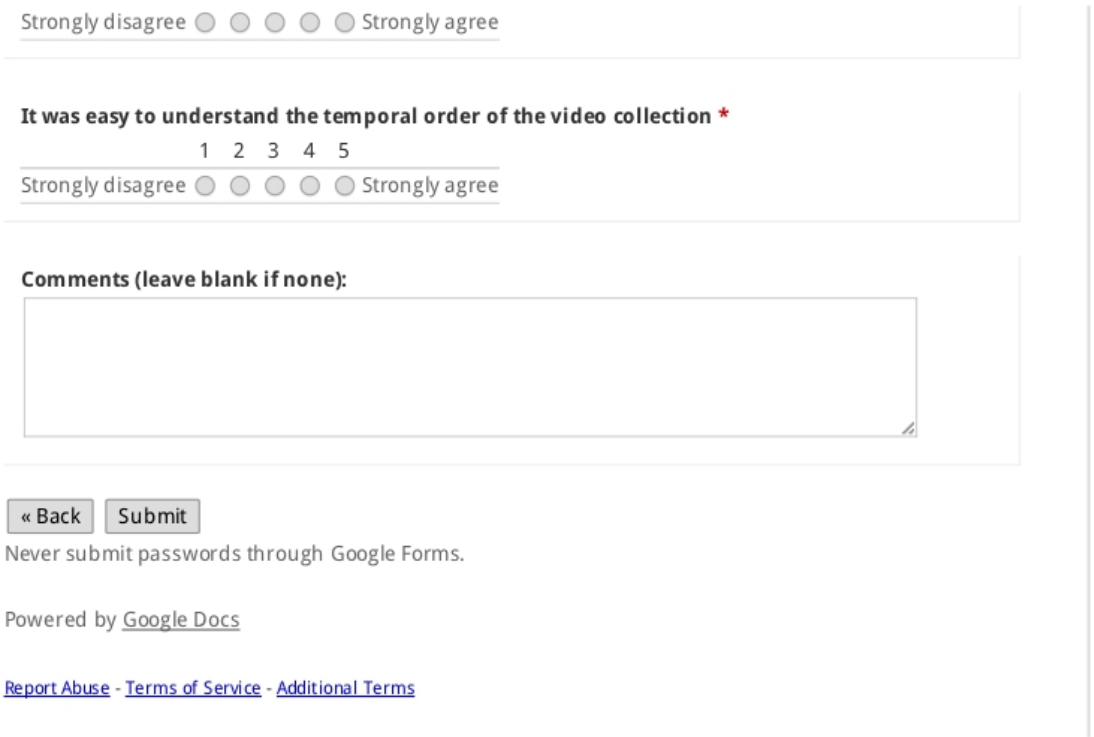


Figure F.12: *Task-related questionnaire (continued).*

Bibliography

[AC76] Michael Argyle and Mark Cook. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, first edition, 1976.

[AFH+97] Anneli Avatare, Emmanuel Frecon, Olof Hagsand, Kai-Mikael Jaa-Aro, Kristian Simsarian, and Olov Stahl. Dive - the distributed interactive virtual environment. Technical report, Swedish Institute of Computer Science, Box 1263, 164 28 Kista, 1997.

[AFS+11] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski. Building rome in a day. *Commun. ACM*, 54(10):105–112, October 2011.

[AHES04] Carlos Hern Andez, Carlos Hern, Ez Esteban, and Francis Schmitt. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding*, 96:367–392, 2004.

[Any02] Anybots, Inc. Anybots. https://www.anybots.com/, 2002. Accessed November 18, 2013.

[App10] Apple, Inc. OpenNi SDK. www.openni.org/downloadfiles, 2010. Last accessed August 05, 2012.

[App14] Apple, Inc. iMovie. http://www.apple.com/mac/imovie/, 2014. Accessed June 10, 2014.

[Arr95] Arrington Research. ViewPoint EyeTracker. http://www.arringtonresearch.com/viewpoint.html, 1995. Accessed August 24, 2012.

[ART96] ART+COM. Timescope. http://www.artcom.de/en/projects/project/detail/timescope/, 1996. Accessed October 30, 2013.

[ART03] ARToolWorks. ARToolkit for iPhone. http://www.artoolworks.com/products/mobile/artoolkit-for-ios/, 2003. Accessed July 15, 2012.

[AW92] Edward H. Adelson and John Y. A. Wang. Single lens stereo with a plenoptic camera. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 14(2):99–106, February 1992.

[AZP+05] Aseem Agarwala, Ke Colin Zheng, Chris Pal, Maneesh Agrawala, Michael Cohen, Brian Curless, David Salesin, and Richard Szeliski. Panoramic video textures. *ACM Transaction on Graphics*, 24(3):821–827, July 2005.

[BBGP10] Mourad Boufarguine, Malek Baklouti, Vincent Guitteny, and Frederic Precioso. Virtu4D: a Real-time Virtualization of Reality. In *3DPVT10*, pages 1–8, 2010.

[BBK07] Christian Beder, Bogumil Bartczak, and Reinhard Koch. A comparison of pmd-cameras and stereo-vision for the task of surface reconstruction using patchlets. In *Proceedings of the Second International ISPRS Workshop*, 2007.

[BBL93] Thomas Baudel and Michel Beaudouin-Lafon. Charade: Remote control of objects using free-hand gestures. *Commun. ACM*, 36(7):28–35, July 1993.

[BBM+01] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 425–432, New York, NY, USA, 2001. ACM.

[BBPP10] Luca Ballan, Gabriel J. Brostow, Jens Puwein, and Marc Pollefeys. Unstructured video-based rendering: interactive exploration of casually captured videos. *ACM Transaction on Graphics*, 29:1–11, July 2010.

[BBRG96] Steve Benford, Chris Brown, Gail Reynard, and Chris Greenhalgh. Shared spaces: transportation, artificiality, and spatiality. In *CSCW '96: Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, pages 77–86, New York, NY, USA, 1996. ACM Press.

[BDR+02] Doug A. Bowman, Ameya Datey, Young Sam Ryu, Umer Farooq, and Omar Vasnaik. Empirical comparison of human behavior and performance with different display devices for virtual environments. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, 2002.

[BFI95] Woodrow Barfield and Thomas A. Furness III. *Virtual environments and advanced interface design*. Oxford University Press, Inc., New York, NY, USA, 1995.

[BGBS02] Patrick Baudisch, Nathaniel Good, Victoria Bellotti, and Pamela Schraedley. Keeping things in context: a comparative evaluation of focus plus context screens, overviews, and zooming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 259–266, New York, NY, USA, 2002. ACM.

[BGR+98] Steve Benford, Chris Greenhalgh, Gail Reynard, Chris Brown, and Boriana Koleva. Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Transaction on Computer-Human Interaction*, 5(3):185–223, September 1998.

[BGTB12] Filippo Bannò, Paolo Simone Gasparello, Franco Tecchia, and Massimo Bergamasco. Real-time compression of depth streams through meshification and valence-based encoding. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '12, pages 263–270, New York, NY, USA, 2012. ACM.

[BIH+12] Alex Butler, Shahram Izadi, Otmar Hilliges, David Molyneaux, Steve Hodges, and David Kim. Shake'n'sense: Reducing structured light interference when multiple depth cameras overlap. In *Proceedings of Human Factors in Computing Systems (ACM CHI)*, New York, NY, USA, April 2012. ACM.

[BK01] Ryad Benosman and Sing B. Kang. *Panoramic vision: sensors, theory, and applications*. Springer-Verlag New York, 2001.

[BL03] Matthew Brown and David G. Lowe. Recognising panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1218–1226. IEEE Computer Society, 2003.

[BL05] Matthew Brown and David G. Lowe. Unsupervised 3d object recognition and reconstruction in unordered datasets. In *Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling*, 3DIM '05, pages 56–63, Washington, DC, USA, 2005. IEEE Computer Society.

[BL07] Matthew Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vision*, 74(1):59–73, August 2007.

[BRB10] J. Birnholtz, A. Ranjan, and R. Balakrishnan. Providing dynamic visual information for collaborative tasks: Experiments with automatic camera control. In *Human-Computer Interaction*, volume 25, pages 261–287, 2010.

[BRB+11] Kai Berger, Kai Ruhl, Christian Brümmer, Yannic Schröder, Alexander Scholz, and Marcus Magnor. Markerless motion capture using multiple color-depth sensors. In *Proc. Vision, Modeling and Visualization (VMV) 2011*, pages 317–324, October 2011.

[Bri10] BrightCom. Brightcom. http://www.brightcom.com/, 2010. Last accessed August 24, 2012.

[Bro96] John Brooke. SUS: A quick and dirty usability scale. In P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland, editors, *Usability Evaluation in Industry*. Taylor and Francis, London, 1996.

[BWB08] Hrvoje Benko, Andrew D. Wilson, and Ravin Balakrishnan. Sphere: multi-touch interactions on a spherical display. In *Proc. the 21st annual ACM symposium on User interface software and technology*, pages 77–86, New York, NY, USA, 2008. ACM.

[BZD+05] Azzedine Boukerche, Anis Zarrad, Diego Duarte, Regina Araujo, and Leonardo Andrade. A novel solution for the development of collaborative virtual environment simulations in large scale. In *Proceedings of the 9th IEEE International Symposium on Distributed Simulation and Real-Time Applications*, DS-RT '05, pages 86–96, Washington, DC, USA, 2005. IEEE Computer Society.

[Cas93] Edward S. Casey. *Getting Back Into Place: Toward a Renewed Understanding of the Place-world*. Indiana University Press, 1993.

[CBMW91] Steve Cook, Gary Birch, Alan Murphy, and John Woolsey. Modelling groupware in the electronic office. *Int. J. Man-Mach. Stud.*, 34(3):369–393, February 1991.

[CGNR08] Jonas Callmer, Karl Granström, J. Nieto, and F. Ramos. Tree of words for visual loop closure detection in urban slam. In Jonghyuk Kim & Robert Mahony, editor, *Proceedings of the 2008 Australasian Conference on Robotics & Automation*, pages 1–8, December 2008.

[Che95] Shenchang Eric Chen. Quicktime vr: An image-based approach to virtual environment navigation. In *Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, pages 29–38, New York, NY, USA, 1995. ACM.

[Che02] Milton Chen. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 49–56. ACM, 2002.

[Cis06] Cisco. Cisco Telepresence. http://www.cisco.com, 2006. Accessed July 16, 2012.

[CKB09] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv.*, 41(1):2:1–2:31, January 2009.

[CKV+09] Krisada Chaiyasarn, Tae-Kyun Kim, Fabio Viola, Roberto Cipolla, and Kenichi Soga. Image mosaicing via quadric surface estimation with priors for tunnel inspection. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 537–540, 2009.

[Cla78] Peter E. Clay. Surrogate travel via optical videodisc. Master's thesis, Massachusetts Institute of Technology, 1978.

[CNSD93] Carolina Cruz-Neira, Daniel J. Sandin, and Thomas A. DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the cave. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 135–142, New York, NY, USA, 1993. ACM.

[Con10] BEAMING Consortium. Beaming website. http://www.beaming-eu.org, 2010. Last accessed July 28, 2013.

[CRG+02] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz, Ivan Tashev, Li-wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, and Steve Silverberg. Distributed meetings: a meeting capture and broadcasting system. In *Proceedings of the ACM International Conference on Multimedia*, pages 503–512, 2002.

[CRZ00] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision*, 40:123–148, November 2000.

[CSC+10] Yan Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3D shape scanning with a time-of-flight camera. In *Conference on Computer Vision and Pattern Recognition (CVPR), 2010 IEEE*, pages 1173 –1180, june 2010.

[CSE06] CSEM. Swissranger. http://www.csem.ch/, 2006. Accessed August 02, 2012.

[CSSH04] Bing-Bing Chai, Sriram Sethuraman, Harpreet S. Sawhney, and Paul Hatrack. Depth map compression for real-time view-based rendering. *Pattern Recognition Letters*, 25:755–766, May 2004.

[CW93] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 279–288, New York, NY, USA, 1993. ACM.

[CZ98] Geoffrey Cross and Andrew Zisserman. Quadric reconstruction from dual-space geometry. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 25–, Washington, DC, USA, 1998. IEEE Computer Society.

[d'A07] Pablo d'Angelo. Hugin - Panorama photo stitcher. http://hugin.sourceforge.net/, 2007. Accessed October 30, 2013.

[des10] VIDERE design. VIDERE design. http://users.rcn.com/mclaughl.dnai/products.htm, 2010. Last accessed Novemeber 2013.

[dHSdVP09] Gerwin de Haan, Josef Scheuer, Raymond de Vries, and Frits H. Post. Egocentric navigation for video surveillance in 3d virtual environments. In *3DUI*, pages 103–110, 2009.

[DKU98] John V. Draper, David B. Kaber, and John M. Usher. Telepresence. *Human Factors*, 40(3):354–375, 1998.

[DLD12] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. *Computer Graphics Forum*, 31(2pt1):305–314, May 2012.

[DLH12] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2018–2025, 2012.

[DLN06] Erick Delage, Honglak Lee, and Andrew Ng. A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image. In *Conference on Computer Vision and Pattern Recognition, 2006 IEEE Computer Society*, volume 2, pages 2418 – 2428, 2006.

[DLN07] Erick Delage, Honglak Lee, and Andrew Ng. Automatic Single-Image 3D reconstructions of indoor manhattan world scenes. In Sebastian Thrun, Rodney Brooks, and Hugh Durrant-Whyte, editors, *Robotics Research*, volume 28 of *Springer Tracts in Advanced Robotics*, chapter 28, pages 305–321. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[DSAO+97] Gwyneth Doherty-Sneddon, Anne Anderson, Claire O'Malley, Steve Langton, Simon Garrod, and Vicki Bruce. Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3(2):105–125, 1997.

[DSAP12] Kevin Dale, Eli Shechtman, Shai Avidan, and Hanspeter Pfister. Multi-video browsing and summarization. In *CVPR Workshops*, pages 1–8, 2012.

[DSKR10] Philip DeCamp, George Shaw, Rony Kubat, and Deb Roy. An immersive system for browsing and visualizing surveillance video. In *ACM Multimedia*, pages 371–380, 2010.

[DTM96] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 11–20, New York, NY, USA, 1996. ACM.

[Duc86] Steve Duck. *Human Relationships: An Introduction to Social Psychology*. Sage Publications, 1986.

[DWH08] Stephen Diverdi, Jason Withert, and Tobias Hllerert. Envisor: Online environment map construction for mixed reality. In *Proceedings of IEEE VR Conference*, 2008.

[DYB98] Paul Debevec, Yizhou Yu, and George Boshokov. Efficient view-dependent image-based rendering with projective texture-mapping. Technical report, University of California at Berkeley Berkeley, Berkeley, CA, USA, 1998.

[Dye01] Charles R. Dyer. Volumetric scene reconstruction from multiple views. In *Foundations of Image Understanding*, pages 469–489. Kluwer, 2001.

[DYNM06] Nicolas Ducheneaut, Nicholas Yee, Eric Nickell, and Robert J. Moore. Alone together?: exploring the social dynamics of massively multiplayer online games. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '06, pages 407–416, New York, NY, USA, 2006. ACM.

[EEH+11] Nikolas Engelhard, Felix Endres, Jurgen Hess, Jurgen Sturm, and Wolfram Burgard. Real-time 3D visual slam with a hand-held rgb-d camera. In *Proceeding of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum*, Vasteras, Sweden, April 2011.

[EHE+12] Felix Endres, Jurgen Hess, Nikolas Engelhard, Jurgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the RGB-D SLAM system. In *Proceeding of the IEEE International Conference on Robotics and Automation (ICRA)*, St. Paul, MA, USA, May 2012.

[Ess02] EssentialReality. EssentialReality Glove. http://www.essentialreality.com/, 2002. Accessed August 24, 2012.

[Fac12] AG Faceshift. Faceshift. http://www.faceshift.com, 2012. Last accessed November 17, 2013.

[Fah97] Lennart Fahln. Distributed interactive virtual environment. http://www.ercim.eu/publication/Ercim_News/enw31/fahlen2.html, 1997. Last accessed Novemeber 2014.

[Far05] Dirk Sven Farin. *Automatic Video Segmentation employing Object/Camera Modeling Techniques*. PhD thesis, Technische Universiteit Eindhoven, 2005.

[Far14] Jason Farman. *Map Interfaces and the Production of Locative Media Space*. Routledge, New York, 2014.

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commununication ACM*, 24(6):381–395, 1981.

[FCSK02] Christoph Fehn, Eddie Cooke, Oliver Schreer, and Peter Kauff. 3D analysis and image-based rendering for immersive tv applications. *Signal Processing: Image Communication*, 17(9):705–715, 2002.

[Feh04] Christoph Fehn. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In A. J. Woods, J. O. Merritt, S. A. Benton, & M. T. Bolas, editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5291 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 93–104, May 2004.

[FER07] Philippe Fechteler, Peter Eisert, and Jurgen Rurainsky. Fast and high resolution 3D face scanning. In *ICIP07*, volume 3, pages 81–84, 2007.

[FFGG+10] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building Rome on a Cloudless Day. In *European Conference on Computer Vision - ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, chapter 27, pages 368–381. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010.

[FFm09] FFmpeg project. FFmpeg. http://www.ffmpeg.org, 2009. Accessed August 06, 2012.

[FGR04] M. Fiala, D. Green, and G. Roth. A panoramic video and acoustic beamforming sensor for videoconferencing. In *Haptic, Audio and Visual Environments and Their Applications*, pages 47–52, October 2004.

[Fil82] Charles J. Fillmore. Towards a descriptive framework for spatial deixis. *Speech, Place and Action: Studies in deixis and related topics*, pages 31–59, 1982.

[FKS00] Susan R. Fussell, Robert E. Kraut, and Jane Siegel. Coordination of communication: Effects of shared visual context on collaborative work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 21–30, New York, NY, USA, 2000. ACM.

[Fli02] Flickr. Flickr image database. http://www.flickr.com/, 2002. Accessed August 02, 2012.

[FM08] Stefan Fuchs and Stefan May. Calibration and registration for precise surface reconstruction with Time-Of-Flight cameras. *International Journal of Intelligent System Technologies Applied*, 5(3/4):274–284, 2008.

[FP05] Yasutaka Furukawa and Jean Ponce. Carved visual hulls for high-accuracy image-based modeling. In *ACM SIGGRAPH 2005 Sketches*, SIGGRAPH '05, New York, NY, USA, 2005. ACM.

[FP07] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, pages 1–8, 2007.

[FPKH09] J. Feulner, J. Penne, E.N.K. Kollorz, and J. Hornegger. Robust real-time 3D modeling of static scenes using solely a time-of-flight sensor. In *OTCBVS09*, pages 74–81, 2009.

[Fre03] Emmanuel Frecon. DIVE: a generic tool for the deployment of shared virtual environments. In *Proceeding of the IEEE Conference on Telecommunications*, volume 1, pages 345–352, June 2003.

[FWSB07] Clifton Forlines, Daniel Wigdor, Chia Shen, and Ravin Balakrishnan. Direct-touch vs. mouse input for tabletop displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 647–656, New York, NY, USA, 2007. ACM.

[GB95] Chris Greenhalgh and Steven Benford. Massive: a collaborative virtual environment for teleconferencing. *ACM Transaction on Computer-Human Interaction*, 2:239–261, September 1995.

[GCS06] Michael Goesele, Brian Curless, and Steven M. Seitz. Multi-view stereo revisited. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2 of *CVPR '06*, pages 2402–2409, Washington, DC, USA, 2006. IEEE Computer Society.

[GGSC96] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 43–54, New York, NY, USA, 1996. ACM.

[GH13] A. Grange and Alvestrand H. A VP9 bitstream overview. Technical report, Google Inc., 2013.

[GJMSMCMJ14] S. Garrido-Jurado, R. Muoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marn-Jimnez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280 – 2292, 2014.

[GL10] Todor Georgiev and Andrew Lumsdaine. Focused plenoptic camera and rendering. *Journal of Electronic Imaging*, 19(2):1–8, 2010.

[Glo06] Global Imagination®. Magic Planet Display. http://www.globalimagination.com/, 2006. Accessed September, 2012.

[Goo01] Google, Inc. Google image database. http://images.google.com/, 2001. Accessed August 02, 2012.

[Goo07] Google, Inc. Google Street View. www.google.com/streetview, 2007. Accessed September 30, 2012.

[Goo08a] Google, Inc. Android Developers. http://developer.android.com/index.html, 2008. Accessed December 13, 2013.

[Goo08b] Google, Inc. Panoramio Look Around. http://www.panoramio.com/, 2008. Accessed October 30, 2013.

[Goo11] Google, Inc. VP8 Data Format and Decoding Guide. http://www.webmproject.org/, 2011. Last accessed Ausugt 06, 2012.

[Goo12] Google. YouTube. https://www.youtube.com/, 2012. Accessed January 15, 2014.

[GP07] Markus Gross and Hanspeter Pfister. *Point-based graphics*. The Morgan Kaufmann Series in Computer Graphics. Morgan Kaufmann, 2007.

[GS10] Marco Gillies and Bernhard Spanlang. Comparing and evaluating real time character engines for virtual environments. *Presence: Teleoper. Virtual Environ.*, 19(2):95–117, April 2010.

[GSC+07] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-View Stereo for Community Photo Collections. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.*, pages 1–8, 2007.

[GSW11] Andreas Girgensohn, Frank Shipman, and Lynn Wilcox. Adaptive clustering and interactive visualizations to support the selection of video clips. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 34:1–34:8, New York, NY, USA, 2011. ACM.

[GWN+03] Markus Gross, Stephan Würmlin, Martin Naef, Edouard Lamboray, Christian Spagno, Andreas Kunz, Esther Koller-Meier, Tomas Svoboda, Luc Van Gool, Silke Lang, Kai Strehlke, Andrew Vande Moere, and Oliver Staadt. Blue-c: a spatially immersive display and 3D video portal for telepresence. *ACM Transaction on Graphics*, 22:819–827, July 2003.

[GYB04] Burak Gokturk, Hakan Yalcin, and Cyrus Bamji. A time-of-flight depth sensor: System description, issues and solutions. In *Sensor3D04*, pages 35–43, 2004.

[HEH05a] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 577–584, New York, NY, USA, 2005. ACM.

[HEH05b] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, volume 1, pages 654–661, Washington, DC, USA, 2005. IEEE Computer Society.

[Hei42] Robert A. Heinlein. *Waldo & Magic, Inc.* Doubleday, New York City, 1942.

[HHR01] Olaf Hall-Holt and Szymon Rusinkiewicz. Stripe boundary codes for real-time structured-light range scanning of moving objects. In *Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001.*, volume 2, pages 359–366, 2001.

[HJS08] Benjamin Huhle, Philipp Jenke, and Wolfgang Strasser. On-the-fly scene acquisition with a handy multi-sensor system. *International Journal of Intelligent System Technologies Applied*, 5:255–263, November 2008.

[HK06] Alexander Hornung and Leif Kobbelt. Robust and efficient photoconsistency estimation for volumetric 3D reconstruction. In *ECCV*, volume 3952 of *LNCS*, pages 179–190, 2006.

[HKH+10] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgbd mapping: Using depth cameras for dense 3D modeling of indoor environments. In *RGB-D: Advanced Reasoning with Depth Cameras Workshop in conjunction with RSS*, 2010.

[HKH12] Daniel Herrera, Juho Kannala, and Janne Heikkilä. Joint depth and color camera calibration with distortion correction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):2058–2064, 2012.

[HL91] Christian Heath and Paul Luff. Collaborative activity and technological design: Task coordination in london underground control rooms. In *Proceedings of the Second Conference on European Conference on Computer-Supported Cooperative Work*, ECSCW'91, pages 65–80, Norwell, MA, USA, 1991. Kluwer Academic Publishers.

[HN06] K Ho and P Newman. Loop closure detection in SLAM by combining visual and spatial appearance. *Robotics and Autonomous Systems*, 54(9):740–749, 2006.

[Ho07] Kin Leon Ho. *Loop closing detection in SLAM using scene appearance*. PhD in Robotics, University of Oxford, Robotics Research Group,Department of Engineering Science, University of Oxford, 2007.

[HRBC06] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. Spatiality in videoconferencing: trade-offs between efficiency and social presence. In *Proceedings of the 2006 20th anniversary conference on Computer Supported Cooperative Work*, CSCW '06, pages 413–422, New York, NY, USA, 2006. ACM.

[HRS92] John A. Hughes, David Randall, and Dan Shapiro. Faltering from ethnography to design. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work*, CSCW '92, pages 115–122, New York, NY, USA, 1992. ACM.

[HSJS10] Benjamin Huhle, Timo Schairer, Philipp Jenke, and Wolfgang Straíer. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Computer Vision and Image Understanding*, 114:1336–1345, December 2010.

[HTC12] HTC Corporation. HTC OneX. http://www.htc.com/www/smartphones/htc-one-x/, 2012. Accessed December 13, 2013.

[HVM+08] Chris Hermans, Cedric Vanaken, Tom Mertens, Frank Van Reeth, and Philippe Bekaert. Augmented panoramic video. *Comput. Graph. Forum*, 27(2):281–290, 2008.

[HYN05] Jinhui Hu, S You, and Ulrich Neumann. Texture painting from video. *Journal of WSCG*, 13(90089):119–125, 2005.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. ISBN: 0521540518.

[IBM09] IBM Corp. IBM SPSS Statistics. http://www-01.ibm.com/software/analytics/spss/products/statistics/, 2009. Accessed December 02, 2013.

[ICWG12] IETF's Codec Working Group. Opus interactive audio codec. http://www.http://opus-codec.org/, 2012. Last accessed November 26, 2013.

[IKG93] Hiroshi Ishii, Minoru Kobayashi, and Jonathan Grudin. Integration of interpersonal space and shared workspace: Clearboard design and experiments. *ACM Transaction on Information System*, 11:349–375, October 1993.

[IKH+11] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 559–568, New York, NY, USA, 2011. ACM.

[IMA13] IMAX Corporation. Imax. https://www.imax.com/about/, 2013. Accessed January 22, 2014.

[IMYV07] Serdar Ince, Emin Martinian, Sehoon Yea, and Anthony Vetro. Depth estimation for view synthesis in multiview video coding. *3DTV Conference (3DTV-CON)*, 1:1–4, 2007.

[Int96] InterSense. InterSense IS900. http://www.intersense.com/pages/20/14, 1996. Accessed August 24, 2012.

[IT93] Ellen A. Isaacs and John C. Tang. What video can and can't do for collaboration: a case study. In *Proceedings of the first ACM International Conference on Multimedia*, MULTIMEDIA '93, pages 199–206, New York, NY, USA, 1993. ACM.

[Jar83] Ray A. Jarvis. A laser time-of-flight range scanner for robotic vision. *PAMI*, 5(5):505–512, September 1983.

[JKC12] Neel Joshi, Abhishek Kar, and Michael Cohen. Looking at You: Fused Gyro and Face Tracking for Viewing Large Imagery on Mobile Devices. In *Proc. SIGCHI '12*, pages 2211–2220, New York, NY, USA, 2012. ACM.

[Jog10] JogAmp. JOGL: Java Binding for the OpenGL API. http://jogamp.org/jogl/www/, 2010. Accessed December 13, 2013.

[JSY03] Hailin Jin, Stefano Soatto, and Anthony J. Yezzi. Multi-view stereo beyond lambert. In *International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 171–178, 2003.

[Jud82] Pearl Judea. Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the American Association of Artificial Intelligence National Conference on AI*, pages 133–136, Pittsburgh, PA, 1982.

[JVS13] Nicole Jochems, Sebastian Vetter, and Christopher Schlick. A comparative study of information input devices for aging computer users. *Behaviour & Information Technology*, 32(9):902–919, 2013.

[KAF+07] Peter Kauff, Nicole Atzpadin, Christoph Fehn, M Müller, Oliver Schreer, Aljosha Smolic, and R Tanger. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. *Signal Processing: Image Communication*, 22(2):217–234, 2007. Special issue on three-dimensional video and television.

[KB04] Leif Kobbelt and Mario Botsch. A survey of point-based techniques in computer graphics. *Computer Graphics Forum*, 28(6):801–814, December 2004.

[KBH06] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics Symposium on Geometry Processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[KBR+12] Julius Kammerl, Nico Blodow, Radu Bogdan Rusu, Suat Gedikli, Michael Beetz, and Eckehard Steinbach. Real-time compression of point cloud streams. In *IEEE International Conference on Robotics and Automation (ICRA)*, Minnesota, USA, May 2012.

[KE12] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[Khr00] Khronos Group. Opengl. http://www.khronos.org/opengl, 2000. Last accessed November 26, 2013.

[Khr03] Khronos Group. Opengl es. http://www.khronos.org/opengles/, 2003. Last accessed November 26, 2013.

[KKC+06] George Kamberov, Gerda Kamberova, O. Chum, J. Kostkov, T. Pajdla, J. Matas, and R. Sara. 3D geometry from uncalibrated images. In *ISVC 2006*, volume 2, pages 802–813, 2006.

[KKZ03] Junhwan Kim, Vladimir Kolmogorov, and Ramin Zabih. Visual correspondence using energy minimization and mutual information. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2 of *ICCV '03*, pages 1033–1041, Washington, DC, USA, 2003. IEEE Computer Society.

[Kni13] Rob Knies. Collaboration, expertise produce enhanced sensing in Xbox One. http://blogs.technet.com/b/microsoft_blog/archive/2013/10/02/collaboration-expertise-produce-enhanced-sensing-in-xbox-one.aspx, 2013. Accessed October, 31, 2013.

[KOLE11] Kihwan Kim, Sangmin Oh, Jeonggyu Lee, and Irfan A. Essa. Augmenting aerial earth maps with dynamic information from videos. *Virtual Reality*, 15(2-3):185–200, 2011.

[KS06] Klaus-Dieter Kuhnert and Martin Stommel. Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction. In *IROS*, pages 4780–4785. IEEE, 2006.

[KSC01] Sing Bing Kang, R. Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001.*, volume 1, pages I–103 – I–110 vol.1, 2001.

[KSK06] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 15–18, 2006.

[KWO10] Michael Kroepfl, Yonatan Wexler, and Eyal Ofek. Efficiently Locating Photographs in Many Panoramas. In *Proc. SIGSPATIAL '10*, New York, NY, USA, 2010. ACM.

[LCDX09] Yebin Liu, Xun Cao, Qionghai Dai, and Wenli Xu. Continuous depth estimation for multi-view stereo. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:2121–2128, 2009.

[LG93] Tony Lindeberg and Jonas Garding. Shape from texture from a multi-scale perspective. In *Proceeding of the 4th International Conference on Computer Vision*, pages 683–691. IEEE Computer Society Press, 1993.

[LH81] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, September 1981.

[LH96] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 31–42, New York, NY, USA, 1996. ACM.

[LHK+03] Paul Luff, Christian Heath, Hideaki Kuzuoka, Jon Hindmarsh, Keiichi Yamazaki, and Shinya Oyama. Fractured ecologies: Creating environments for collaboration. *Hum.-Comput. Interact.*, 18(1):51–84, June 2003.

[Lim07] Engineered Arts Limited. Robothespian. http://www.robothespian.co.uk, 2007. Accessed November 18, 2013.

[Lin03] Linden Research, Inc. Second life. http://secondlife.com/, 2003. Accessed January 13, 2014.

[Lip80] Andrew Lippman. Movie-maps: An application of the optical videodisc to computer graphics. *SIGGRAPH Comput. Graph.*, 14(3):32–42, July 1980.

[LK06] M. Lindner and A. Kolb. Lateral and depth calibration of pmd-distance sensors. In *Proc. Int. Symp. on Visual Computing*, LNCS, pages 524–533. Springer, 2006.

[LLBM09] Christian Lipski, Christian Linz, Kai Berger, and Marcus Magnor. Virtual video camera: image-based viewpoint navigation through space and time. In *ACM SIGGRAPH '09: Posters*, SIGGRAPH '09, pages 93:1–93:1, New York, NY, USA, 2009. ACM.

[Log03] LifeSize Logitech. Lifesize. http://www.lifesize.com/, 2003. Last accessed August 24, 2012.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LP99] Jochen Lang and Dinesh K. Pai. Bayesian estimation of distance and surface normal with a time-of-flight laser rangefinder. In *Proceedings of the 2Nd International Conference on 3-D Digital Imaging and Modeling*, 3DIM'99, pages 109–117, Washington, DC, USA, 1999. IEEE Computer Society.

[LS09] James R. Lewis and Jeff Sauro. The factor structure of the system usability scale. In *HCI (10)*, pages 94–103, 2009.

[LWG04] Edouard Lamboray, Stephan Würmlin, and Markus Gross. Real-time streaming of point-based 3D video. In *Proceedings of IEEE Virtual Reality*, pages 91–98. IEEE Computer Society Press, 2004.

[LWG05] Edouard Lamboray, Stephan Würmlin, and Markus Gross. Data streaming in telepresence environments. *IEEE Transactions on Visualization and Computer Graphics*, 11:637–348, 2005.

[LWL+10] Kai Liu, Yongchang Wang, Daniel L. Lau, Qi Hao, and Laurence G. Hassebrook. Dual-frequency pattern scheme for high-speed 3-D shape measurement. *Optics Express*, 18(5):5229–5244, March 2010.

[LYKH11] Paul Luff, Naomi Yamashita, Hideaki Kuzuoka, and Christian Heath. Hands on hitchcock: Embodied reference to a moving scene. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 43–52, New York, NY, USA, 2011. ACM.

[MAW+07] Paul Merrell, Amir Akbarzadeh, Liang Wang, Jan michael Frahm, and Ruigang Yang David Nister. Real-time visibility-based fusion of depth maps. In *Internatinal Conference on Computer Vision and Pattern Recognition*, 2007.

[MBX+06] Emin Martinian, Alexander Behrens, Jun Xin, Anthony Vetro, and Huifang Sun. Extensions of h.264/avc for multiview video compression. In *IEEE International Conference on Image Processing*, 2006.

[McC93] Scott McCloud. *Understanding comics: The invisble art.* Harper Collins Publishers, 1993.

[McC07] Neil james McCurdy. *RealityFlythrough: A system for ubiquitous video*. Ph.d., University of California, San Diego, 2007.

[MCN94] Shawna Meyer, Oryx Cohen, and Erik Nilsen. Device comparisons for goal-directed drawing tasks. In *Conference Companion on Human Factors in Computing Systems*, CHI '94, pages 251–252, New York, NY, USA, 1994. ACM.

[MCRTB10] Betty J. Mohler, Sarah H. Creem-Regehr, William B. Thompson, and Heinrich H. Bülthoff. The effect of viewing a self-avatar on distance judgments in an HMD-based virtual environment. *Presence: Teleoper. Virtual Environ.*, 19(3):230–242, June 2010.

[MD08] Matthieu Maitre and Minh N. Do. Joint encoding of the depth image based representation using shape-adaptive wavelets. In *15th IEEE International Conference on Image Processing, 2008. ICIP 2008.*, pages 1768–1771, October 2008.

[Mee99] Michael Meehan. Survey of multi-user distributed virtual environments. *ACM Transactions on Graphics (SIGGRAPH)*, 99:8–13, 1999.

[MF11] Andrew Maimone and Henry Fuchs. A first look at a telepresence system with room-sized real-time 3D capture and large tracked display. In *The 21st International Conference on Artificial Reality and Telexistence (ICAT)*. 2011.

[MGVL02] Maja Matijasevic, Denis Gracanin, Kimon P. Valavanis, and Ignac Lovrek. A framework for multiuser distributed virtual environments. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE*, 32(4):416–429, August 2002.

[Mic02] Microsoft® Skype Division. Skype. http:skype.com, 2002. Accessed July 25, 2012.

[Mic08] Microsoft®. Photosynth. http://photosynth.net, 2008. Accessed September 02, 2012.

[Mic10] Michael Naimark. Time Binoculars. http://www.naimark.net/projects/pending/timebinoculars.htm, 2010. Accessed October 30, 2013.

[Mic12a] Microsoft®. Kinect™ for Windows. http://www.microsoft.com/en-us/kinectforwindows/, 2012. Accessed August 02, 2012.

[Mic12b] Microsoft®. Windows Live Movie Maker. http://windows.microsoft.com/en-us/windows-live/movie-maker, 2012. Accessed April 02, 2014.

[Mic12c] Microsoft® Research. Image Composite Editor. http://research.microsoft.com/en-us/um/redmond/groups/ivm/ICE/, 2012. Accessed November, 02, 2012.

[Min80] Marvin Minsky. Telepresence. *Omni*, 2(9):45–52, June 1980.

[MJSS02] David W. Mizell, Sthephen P. Jones, Mel Slater, and Bernhard Spanlang. Comparing immersive virtual reality with other display modes for visualizing complex 3d geometry. Technical report, University College London, 2002.

[MMS+09] Philipp Merkle, Yannick Morvan, Aljoscha Smolic, Dirk Farin, Karsten Mueller, Peter H. N. de With, and Thomas Wiegand. The effects of multiview depth video compression on multiview rendering. *Singal Processing: Image Communication*, 24(1-2):73–88, 2009.

[Moe] Andrew Vande Moere. Blue-c visuals. http://blue-c.ethz.ch/index.php?menu=visuals&sub=pictures. Last accessed Novemeber 2014.

[Mol69] Julius P. Molnar. Picturephone service-a new way of communicating. volume 47:5, pages 134–135. Bell Laboratories Record, 1969.

[Mor83] Hans P. Moravec. The Stanford Cart and the CME Rover. *PIEEE*, 71(7):872–884, July 1983.

[MP90] Jitendra Malik and Pietro Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America*, 7:923–932, 1990.

[MR97] Jitendra Malik and Ruth Rosenholtz. Computing local surface orientation and shape from texture forcurved surfaces. *International Journal of Computer Vision*, 23:149–168, June 1997.

[MRWB03] Michael Meehan, Sharif Razzaque, Mary C. Whitton, and Frederick P. Brooks, Jr. Effect of latency on presence in stressful virtual environments. In *Proceedings of the IEEE Virtual Reality 2003*, VR '03, pages 141–, Washington, DC, USA, 2003. IEEE Computer Society.

[MSD+12] Alessandro Mulloni, Hartmut Seichter, Andreas Dünser, Patrick Baudisch, and Dieter Schmalstieg. 360 degrees: panoramic overviews for location-based services. In *Proceedings of the Annual Conference on Human Factors in Computing Systems*, pages 2565–2568. ACM, 2012.

[MSGF99] Aditi Majumder, W. Brent Seales, Meenakshisundaram Gopi, and Henry Fuchs. Immersive teleconferencing: a new algorithm to generate seamless panoramic video imagery. In *Proceedings of the International Conference on Multimedia*, pages 169–178. ACM, 1999.

[MSN05] Jeff Michels, Ashutosh Saxena, and Andrew Y. Ng. High speed obstacle avoidance using monocular vision and reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 593–600, New York, NY, USA, 2005. ACM.

[MTS+05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1):43–72, November 2005.

[MWS06] Detlev Marpe, Thomas Wiegand, and Gary J. Sullivan. The h.264/mpeg4 advanced video coding standard and its applications. *Communications Magazine, IEEE*, 44(8):134–143, August 2006.

[MWW02] Atsuto Maki, Mutsumi Watanabe, and Charles Wiles. Geotensity: Combining motion and lighting for 3d surface reconstruction. In *International Journal of Computer Vision*, volume 48, pages 75–90, 2002.

[MYD+13] Andrew Maimone, Xubo Yang, Nate Dierk, Andrei State, Mingsong Dou, and Henry Fuchs. General-purpose telepresence with head-worn optical see-through displays and projector-based lighting. In *Virtual Reality (VR), 2013 IEEE*, pages 23–26, 2013.

[Nai06] Michael Naimark. Aspen the verb: Musings on heritage and virtuality. *Presence: Teleoper. Virtual Environ.*, 15(3):330–335, June 2006.

[Nat96] NaturalPoint. OptiTrack. http://www.naturalpoint.com/optitrack/, 1996. Accessed August 24, 2012.

[ND10] Richard A. Newcombe and Andrew J. Davison. Live dense reconstruction with a single moving camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*, pages 1498–1505, June 2010.

[New97a] William Newman. The use of critical parameters in the design of web-based interactive systems. *Time and the Web*, 1997.

[New97b] William M. Newman. Better or just different? on the benefits of designing interactive systems in terms of critical parameters. In *Proceedings of the 2Nd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '97, pages 239–245, New York, NY, USA, 1997. ACM.

[New01] New House Internet Services BV. Ptgui. http://www.ptgui.com/, 2001. Accessed July 17, 2012.

[NG13] Hanna Nyqvist and Fredrik Gustafsson. A high-performance tracking system based on camera and imu. In *Information Fusion (FUSION), 2013 16th International Conference on*, pages 2065–2072, 2013.

[NH05] Paul M. Newman and Kin Leong Ho. Slam-loop closing with visually salient features. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005. ICRA*, pages 635–642, April 2005.

[NLB+05] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light Field Photography with a Hand-Held Plenoptic Camera. Technical report, Stanford University Computer Science, April 2005.

[NLD11] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *IEEE International Conference on Computer Vision (ICCV), 2011*, pages 2320–2327, November 2011.

[NMK09] Hideyuki Nakanishi, Yuki Murakami, and Kei Kato. Movable cameras enhance social telepresence in media spaces. In *Proceedings of the 27th international conference on Human Factors in Computing Systems*, CHI '09, pages 433–442, New York, NY, USA, 2009. ACM.

[NRK98] P. J. Narayanan, Peter W. Rander, and Takeo Kanade. Constructing virtual worlds using dense stereo. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 3–11, Washington, DC, USA, 1998. IEEE Computer Society.

[NSQ12] James Norris, Holger Schnädelbach, and Guoping Qiu. Camblend: an object focused collaboration tool. In *Proceedings of SIGCHI '12*, New York, NY, USA, 2012. ACM.

[NVI08] NVIS, Inc. nVisor SX111. http://www.nvisinc.com/product.php?id=48, 2008. Last accessed November 22, 2013.

[NYH+03] Ulrich Neumann, Suya You, Jinhui Hu, Bolan Jiang, and Jong Weon Lee. Augmented Virtual Environments (AVE): Dynamic Fusion of Imagery and 3D Models. In *VR*, pages 61–, 2003.

[Ocu10] Oculus VR, Inc. Raknet. http://www.jenkinssoftware.com/, 2010. Last accessed July 28, 2012.

[Ocu12] Oculus VR, Inc. Oculus rift - virtual reality headset for 3d gaming. http://www.oculusvr.com/, 2012. Accessed December 28, 2013.

[Ogg10] Thierry Oggier. Time of flight principle diagram. http://commons.wikimedia.org/wiki/File:TOF-camera-principle.jpg, 2010. Accessed January 22, 2014.

[OLA+96] Claire O'Malley, Steve Langton, Anne Anderson, Gwyneth Doherty-Sneddon, and Vicki Bruce. Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, 8(2):177–192, 1996.

[Oli99] John Oliensis. A multi-frame structure-from-motion algorithm under perspective projection. *International Journal of Computer Vision*, 34:163–192, October 1999.

[Ope06] OpenFrameworks. OpenFrameworks. http://www.openframeworks.cc, 2006. Accessed July 15, 2012.

[OSS12] Oyewole Oyekoya, William Steptoe, and Anthony Steed. Sphereavatar: a situated display to represent remote collaborator. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 2551–2560, New York, NY, USA, 2012. ACM.

[OSS+13] Oyewole Oyekoya, Ran Stone, William Steptoe, Laith Alkurdi, Stefan Klare, Angelika Peer, Tim Weyrich, Benjamin Cohen, Franco Tecchia, and Anthony Steed. Supporting interoperability and presence awareness in collaborative mixed reality environments. In *Proc. of the 19th ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 165–174, October 2013.

[PCD+12] Sören Pirk, Michael F. Cohen, Oliver Deussen, Matt Uyttendaele, and Johannes Kopf. Video enhanced gigapixel panoramas. In *SIGGRAPH Asia 2012 Technical Briefs*, SA '12, pages 7:1–7:4, New York, NY, USA, 2012. ACM.

[PCS+00] Emilee Patrick, Dennis Cosgrove, Aleksandra Slavkovic, Jennifer A. Rode, Thom Verratti, and Greg Chiselko. Using a large projection screen as an alternative to head-mounted displays for virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '00, pages 478–485, New York, NY, USA, 2000. ACM.

[PHE+11] Dawid Pajak, Robert Herzog, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. Scalable remote rendering with depth and motion-flow augmented streaming. *Computer Graphics Forum*, 30(2):415–424, 2011. Proceedings Eurographics 2011.

[PKB05] Nicholas F. Polys, Seonho Kim, and Doug A. Bowman. Effects of information layout, screen size, and field of view on user performance. In *in Information-Rich Virtual Environments, Proceedings of ACM Symposium on Virtual Reality Software and Technology*, pages 46–55, 2005.

[PKVG98] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV '98)*, Washington, DC, USA, 1998. IEEE Computer Society.

[PKVVG98] Marc Pollefeys, Reinhard Koch, Maarten Vergauwen, and Luc Van Gool. Metric 3D surface reconstruction from uncalibrated image sequences. In Reinhard Koch and Luc Van Gool, editors, *3D Structure from Multiple Images of Large-Scale Environments*, volume 1506 of *Lecture Notes in Computer Science*, pages 139–154. Springer Berlin / Heidelberg, 1998.

[PKW11] Fabrizio Pece, Jan Kautz, and Tim Weyrich. Adapting standard video codecs for depth streaming. In *Proceedings of the 17th Eurographics Conference on Virtual Environments & Third Joint Virtual Reality*, EGVE - JVRC'11, pages 59–66, Aire-la-Ville, Switzerland, Switzerland, 2011. Eurographics Association.

[PMD09] PMD[vision]®. CamCube3.0. http://www.pmdtec.com/, 2009. Accessed August 02, 2012.

[Poi10a] Point Grey Research. Bumblebee XB3. http://ww2.ptgrey.com/stereo-vision/bumblebee-xb3, 2010. Accessed January 22, 2013.

[Poi10b] Point Grey Research. LadyBug3. http://www.ptgrey.com/products/ladybug3/Ladybug3_360_video_camera.asp, 2010. Accessed January 22, 2013.

[Poi11] Point Grey Research. Overview of the Ladybug image stitching process. http://www.ptgrey.com/KB/10564, January 2011. Accessed December 2014.

[Pol00] Polhemus. Fastrack. http://www.polhemus.com/?page=motion_fastrak, 2000. Accessed August 24, 2012.

[Pol10] Polycom®. CX500. http://www.polycom.com, 2010. Accessed July 17, 2012.

[Pol11] Polycom®. RealPresence Immersive Studio. http://www.polycom.com/products-services/hd-telepresence-video-conferencing/

realpresence-immersive/realpresence-immersive-studio.html, 2011. Accessed July, 2014.

[Por03] F. Porikli. Inter-camera color calibration by correlation model function. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–133–6 vol.3, Sept 2003.

[Por06] PortAudio. Portaudio. http://www.portaudio.com/, 2006. Last accessed September 01, 2012.

[PRI+13] Vivek Pradeep, Christoph Rhemann, Shahram Izadi, Christopher Zach, Michael Bleyer, and Steven Bathiche. Monofusion: Real-time 3d reconstruction of small scenes with a single web camera. In *ISMAR*, pages 83–88, 2013.

[PS07] Xueni Pan and Mel Slater. A Preliminary Study of Shy Males Interacting with a Virtual Female. In *PRESENCE 2007: The 10th Annual International Workshop on Presence*, 2007.

[PSP93] Randy Pausch, M. Anne Shackelford, and Dennis Proffitt. A user study comparing head-mounted and stationary displays. In *Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in*, pages 41–45, 1993.

[PSW+13] Fabrizio Pece, William Steptoe, Fabian Wanner, Simon Julier, Tim Weyrich, Jan Kautz, and Anthony Steed. Panoinserts: mobile spatial teleconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1319–1328, New York, NY, USA, 2013. ACM.

[PTP+14] Fabrizio Pece, James Tompkin, Hanspeter Pfister, Jan Kautz, and Christian Theobalt. Device Effect on Panoramic Video+Context Tasks. In *Proceedings of the 2014 Conference on Visual Media Production*, CVMP '14, New York, NY, USA, 2014. ACM.

[PVG02] Marc Pollefeys and Luc Van Gool. From images to 3D models. In *Communications of the ACM*, volume 45, pages 50–55, July 2002.

[PVGV+04] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59:207–232, September 2004.

[PWC08] Suporn Pongnumkul, Jue Wang, and Michael F. Cohen. Creating map-based storyboards for browsing tour videos. In *UIST*, pages 13–22, 2008.

[QT11] Inc. Qualcomm Technologies. Qualcomm Vuforia. http://www.qualcomm.com/solutions/augmented-reality, 2011. Accessed November 26, 2013.

[RBB06] Abhishek Ranjan, Jeremy P. Birnholtz, and Ravin Balakrishnan. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. In

Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW '06, pages 403–412, New York, NY, USA, 2006. ACM.

[RD06] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, volume 1, pages 430–443, May 2006.

[RdBF+02] Andre Redert, Marc Op de Beeck, Christoph Fehn, Wijnand IJsselsteijn, Marc Pollefeys, Luc Van Gool, Eyal Ofek, Ian Sexton, and Philip Surman. Attest: Advanced three-dimensional television system technologies. *3D Data Processing Visualization and Transmission, International Symposium on*, 0:313–321, 2002.

[RDP+11] Malcolm Reynolds, Jozef Doboš, Leto Peel, Tim Weyrich, and Gabriel J. Brostow. Capturing time-of-flight data with confidence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 945–952, 2011.

[RFC97] Charlie Rothwell, Olivier D. Faugeras, and Gabriella Csurka. A comparison of projective reconstruction methods for pairs of views. *Computer Vision and Image Understanding*, 68(1):37–58, October 1997.

[RGC01] Yong Rui, Anoop Gupta, and J.J. Cadiz. Viewing meeting captured by an omnidirectional camera. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 450–457. ACM, 2001.

[RHQ14] Katja Rogers, Uta Hinrichs, and Aaron Quigley. It doesn't compare to being there: In-situ vs. remote exploration of museum collections. In *Workshop The Search Is Over! Exploring Cultural Collections with Visualization, held in conjunction with the Digital Libraries Conference (DL'14)*, 2014.

[RL01] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *Third International Conference on 3-D Digital Imaging and Modeling, 2001. Proceedings.*, pages 145–152, 2001.

[Rog83] G. F. C Rogers. *The Nature of Engineering: a Philosophy of Technology.* London: Macmillan, 1983.

[RSD+12] Christian Richardt, Carsten Stoll, Neil A. Dodgson, Hans-Peter Seidel, and Christian Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. *Computer Graphics Forum*, 31(2pt1):247–256, May 2012.

[RSPB05] Bernhard E Riecke, Joerg Schulte-Pelkum, and Heinrich H Buelthoff. Perceiving simulated ego-motions in virtual reality: Comparing large screen displays with HMDs. In *Electronic Imaging 2005*, pages 344–355. International Society for Optics and Photonics, 2005.

[RWC+98] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th annual conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pages 179–188, New York, NY, USA, 1998. ACM.

[SA91] Minas Spetsakis and John Yiannis Aloimonos. A multi-frame approach to visual motion perception. *International Journal of Computer Vision*, 6:245–255, August 1991.

[Sam11] Samsung Electronics Co., LTD. Samsung Galaxy SII. om/global/microsite/galaxys2/html/, 2011. Accessed December 13, 2013.

[SB11] Klaus Schoeffmann and Laszlo Boeszoermenyi. Image and video browsing with a cylindrical 3d storyboard. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 63:1–63:2, New York, NY, USA, 2011. ACM.

[SBA92] Abigail Sellen, Bill Buxton, and John Arnott. Using spatial cues to improve videoconferencing. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '92, pages 651–652, New York, NY, USA, 1992. ACM.

[SCD+06] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, Washington, DC, USA, 2006. IEEE Computer Society.

[SCKMB03] Herring Susan C, Borner Katy, and Swan Maggie B. When rich media are opaque: Spatial reference in a 3-d virtual world. *The Journal of Collaborative Computing*, 2003.

[SCMS01] Gregory G. Slabaugh, W. Bruce Culbertson, Thomas Malzbender, and Ronald W. Schafer. A survey of methods for volumetric scene reconstruction from photographs. In *Volume Graphics'01*, pages 81–100, 2001.

[SCN07] Ashutosh Saxena, Sung Chung, and Andrew Y. Ng. 3-D depth reconstruction from a single still image. *International Journal of Computer Vision*, 76:2007, 2007.

[SGSS08] Noah Snavely, Rahul Garg, Steven M. Seitz, and Richard Szeliski. Finding paths through the world's photos. *ACM Transactions on Graphics (SIGGRAPH 2008)*, 27(3):11–21, 2008.

[Shi11] Anand Lal Shimpi. The Sandy Bridge Review. http://tinyurl.com/2u25zjj, 2011. Last accessed December 10, 2013.

[SHYN03] Ismail Oner Sebe, Jinhui Hu, Suya You, and Ulrich Neumann. 3d video surveillance with augmented virtual environments. In *First ACM SIGMM International Workshop on Video Surveillance*, IWVS '03, pages 107–112, New York, NY, USA, 2003. ACM.

[SJF+13] Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey, and Giuliano Maciocci. Bethere: 3d mobile collaboration with spatial input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 179–188, New York, NY, USA, 2013. ACM.

[SJP11] Jan Smisek, Michal Jancosek, and Tomas Pajdla. 3D with Kinect. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1154–1160, Nov 2011.

[SK93] Richard Szeliski and Sing Bing Kang. Recovering 3D shape and motion from image streams using nonlinear least squares. In *1993 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 1993)*, pages 752 –753, jun 1993.

[SK05] Aljosha Smolic and Peter Kauff. Interactive 3-D video representation and coding technologies. *Proceedings of the IEEE*, 93(1):98–110, January 2005.

[Sla09] Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos Trans R Soc Lond B Biol Sci*, 364(1535):3549–3557, Dec 2009.

[SLK11] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 667–674, Washington, DC, USA, 2011. IEEE Computer Society.

[SLK12] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, ECCV'12, pages 752–765, Berlin, Heidelberg, 2012. Springer-Verlag.

[SLP+07] Agnes Swadzba, Bing Liu, Jochen Penne, Oliver Jesorsky, and Ralf Kompe. A comprehensive system for 3D modeling from range images acquired from a 3D tof sensor. In *International Conference on Computer Vision Systems*, Bielefeld University, Bielefeld, Germany, March 2007. University Library of Bielefeld, University Library of Bielefeld.

[SLU+96] Mel Slater, Vasilis Linakis, Martin Usoh, Rob Kooper, and Gower Street. Immersion, presence, and performance in virtual environments: An experiment with tridimensional chess. In *ACM Virtual Reality Software and Technology (VRST)*, pages 163–172, 1996.

[SLW07] Simon T. Suen, Edmund Y. Lam, and Kenneth K. Wong. Photographic stitching with optimized object and color matching based on image derivatives. *Optics Express*, 15(12):7689–7696, 2007.

[SM04] R. William Soukoreff and I. Scott MacKenzie. Towards a standard for pointing device evaluation, perspectives on 27 years of fitts' law research in hci. *Int. J. Hum.-Comput. Stud.*, 61(6):751–789, December 2004.

[SMS09] Farzan Sasangohar, I Scott MacKenzie, and Stacey D Scott. Evaluation of mouse and touch input for a tabletop display using fitts' reciprocal tapping task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 12:53, pages 839–843. SAGE Publications, 2009.

[SNO+12] William Steptoe, Jean-Marie Normand, Oyewole Oyekoya, Fabrizio Pece, Elias Giannopoulos, Franco Tecchia, Anthony Steed, Tim Weyrich, Jan Kautz, and Mel Slater. Acting rehearsal in collaborative multimodal mixed reality environments. *Presence*, 21(4):406–422, 2012.

[Sof11] SoftKinetic™. Depthsense™. http://www.softkinetic.com/solutions/depthsensecameras.aspx, 2011. Accessed August 02, 2014.

[SOM+09] William Steptoe, Oyewole Oyekoya, Alessio Murgia, Robin Wolff, John Rae, Estefania Guimaraes, David Roberts, and Anthony Steed. Eye tracking for avatar eye gaze control during object-focused multiparty interaction in immersive collaborative virtual environments. *Virtual Reality Conference, IEEE*, 1:83–90, 2009.

[Son12] Hai Tao Song. Updating Fitts' law to account for restricted display field of view conditions. *International Journal of Human-Computer Interaction*, 28(4):269–279, 2012.

[SPS05] Drew Steedly, Chris Pal, and Richard Szeliski. Efficiently registering video into panoramic mosaics. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pages 1300–1307, Washington, DC, USA, 2005. IEEE Computer Society.

[SS91] Andrew Sears and Ben Shneiderman. High precision touchscreens: design strategies and comparisons with a mouse. *Int. J. Man-Mach. Stud.*, 34(4):593–613, April 1991.

[SS97] Richard Szeliski and Heung-Yeung Shum. Creating full view panoramic image mosaics and environment maps. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 251–258. ACM Press/Addison-Wesley Publishing Co., 1997.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, April 2002.

[SS03] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Computer Society Conference On Computer Vision and Pattern Recognition*, pages 195–202, 2003.

[SSA+01] Ralph Schroeder, Anthony Steed, Ann-Sofie Axelsson, Ilona Heldal, sa Abelin, Josef Widestrm, Alexander Nilsson, and Mel Slater. Collaborating in networked immersive spaces: as good as being there together? *Computers and Graphics*, 25(5):781 – 788, 2001.

[SSG00] Scott B. Steinman, Barbara A. Steinman, and Ralph Philip Garzia. *Foundations of Binocular Vision: A Clinical Perspective.* McGraw-Hill Companies, New York, first edition, 2000.

[SSH+03] Anthony Steed, Maria Spante, Ilona Heldal, Ann-Sofie Axelsson, and Ralph Schroeder. Strangers and friends in caves: an exploratory study of collaboration in networked ipt systems for extended periods of time. In *Proceedings of the 2003 symposium on Interactive 3D graphics*, I3D '03, pages 51–54, New York, NY, USA, 2003. ACM.

[SSM11] Jeferson R. Silva, Thiago T. Santos, and Carlos H. Morimoto. Virtual reality in brazil: Automatic camera control in virtual environments augmented using multiple sparse videos. *Comput. Graph.*, 35(2):412–421, April 2011.

[SSN07a] Ashutosh Saxena, Jamie Schulte, and Andrew Y. Ng. Depth estimation using monocular and stereo cues. In *Proceedings of the 20th international joint conference on Artifical Intelligence*, IJCAI'07, pages 2197–2203, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[SSN07b] Ashutosh Saxena, Min Sun, and Andrew Ng. Learning 3-D scene structure from a single still image. In *IEEE 11th International Conference on Computer Vision, 2007 (ICCV 2007)*, pages 1–8, October 2007.

[SSO+12] Anthony Steed, William Steptoe, Wole Oyekoya, Fabrizio Pece, Tim Weyrich, Jan Kautz, Doron Friedman, Angelika Peer, Massimiliano Solazzi, Franco Tecchia, Massimo Bergamasco, and Mel Slater. Beaming: An asymmetric telepresence system. *IEEE Comput. Graph. Appl.*, 32(6):10–17, November 2012.

[SSRR10] William Steptoe, Anthony Steed, Aitor Rovira, and John Rae. Lie tracking: social presence, truth and deception in avatar-mediated telecommunication. In *Proceedings*

of the 28th international conference on Human factors in Computing Systems, CHI '10, pages 1039–1048, New York, NY, USA, 2010. ACM.

[SSS06] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. In *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[SSS08] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from Internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, November 2008.

[Sta11] Starlab. Enobio. http://starlab.es/products/enobio, 2011. Accessed August 24, 2012.

[STDT08] Sebastian Schuon, Christian Theobalt, James Davis, and Sebastian Thrun. High-quality scanning using time-of-flight depth superresolution. In *TOF-CV08*, pages 1–7, 2008.

[STDT09] Sebastian Schuon, Christian Theobalt, James Davis, and Sebastian Thrun. Lidarboost: Depth superresolution for tof 3D shape scanning. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:343–350, 2009.

[Ste96] Anthony James Steed. *Defining Interaction within Immersive Virtual Environments*. Computer science, Department of Computer Science, Queen Mary and Westfield College, University of London, 1996.

[Ste10] William Steptoe. *Eye Tracking and Avatar-Mediated Communication in Immersive Collaborative Virtual Environment*. PhD thesis, Department of Computer Science, University College London, London, 2010.

[SU93] Mel Slater and Martin Usoh. Presence in immersive virtual environments. In *Virtual Reality Annual International Symposium, 1993., 1993 IEEE*, pages 90–96, 1993.

[SUS94] Mel Slater, M Usoh, and Anthony James Steed. Depth of presence in immersive virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 3(2):130–144, 1994. Depth of Presence in Immersive Virtual Environments xD;TY - JOUR.

[Sut65] Ivan E. Sutherland. The ultimate display. In *Proceedings of the IFIP Congress*, pages 506–508, 1965.

[Sut68] Ivan E. Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, AFIPS '68 (Fall, part I), pages 757–764, New York, NY, USA, 1968. ACM.

[Sze94] Richard Szeliski. Image mosaicing for tele-reality applications. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pages 44–53, 1994.

[Sze06] Richard Szeliski. Image alignment and stitching: a tutorial. *Foundations and Trends in Computer Graphics and Computer Vision*, 2:1–104, 2006.

[Sze10] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, 1st edition, 2010.

[TBP10] Aparna Taneja, Luca Ballan, and Marc Pollefeys. Modeling dynamic scenes recorded with freely moving cameras. In *ACCV (3)*, pages 613–626, 2010.

[TCB+10] Franco Tecchia, Marcello Carrozzino, Sandro Bacinelli, Fabio Rossi, Davide Vercelli, Giuseppe Marino, Paolo Simone Gasparello, and Massimo Bergamasco. A flexible framework for wide-spectrum vr development. *Presence*, 19(4):302–312, 2010.

[Tel12] Telanetix, Inc. Telanetix. http://www.telanetix.com/, 2012. Last accessed August 24, 2012.

[TFK+02] Michael Tsang, George W. Fitzmaurice, Gordon Kurtenbach, Azam Khan, and Bill Buxton. Boom chameleon: Simultaneous capture of 3d viewpoint, voice and gesture annotations on a spatially-aware display. In *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology*, UIST '02, pages 111–120, New York, NY, USA, 2002. ACM.

[TG98] Costa Touma and Craig Gotsman. Triangle mesh compression. In *Graphics interface*, volume 98, pages 26–34, 1998.

[TGSP03] Desney S. Tan, Darren Gergle, Peter Scupelli, and Randy Pausch. With similar visual angles, larger displays improve spatial performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 217–224, New York, NY, USA, 2003. ACM.

[TK92] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, November 1992.

[TKBMW12] Maurice Ten Koppel, Gilles Bailly, Jörg Müller, and Robert Walter. Chained displays: configurations of public displays can be used to influence actor-, audience-, and passer-by behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 317–326, New York, NY, USA, 2012. ACM.

[TKKT12] James Tompkin, Kwang In Kim, Jan Kautz, and Christian Theobalt. Videoscapes: exploring sparse, unstructured video collections. *ACM Trans. Graph.*, 31(4):68, 2012.

[TKM+13] Petri Tanskanen, Kalin Kolev, Lorenz Meier, Federico Camposeco, Olivier Saurer, and Marc Pollefeys. Live metric 3d reconstruction on mobile phones. In *ICCV*, pages 65–72, 2013.

[TLF10] E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010.

[TM08] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, July 2008.

[TMHF00] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, 2000. Springer-Verlag.

[TPS+13] James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz, and Christian Theobalt. Video collections in panoramic contexts. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 131–140, New York, NY, USA, 2013. ACM.

[Uni05] Unity Technologies. Unity 3. http://unity3d.com, 2005. Accessed July 18, 2012.

[Urm00] Chris Urmson. A comparison of pt. grey researchs digiclops and videre designs small vision system for sensing on the hyperion robot. Technical report, Carnegie Mellon University, 2000.

[USRO02] A. Ullrich, N. Studnicka, J. Riegl, and S. Orlandini. Long-range high-performance time-of-flight-based 3D imaging sensors. In *3DPVT02*, pages 852–855, 2002.

[Vin91] W. G. Vincenti. *What Engineers Know and How They Know It: Analytical Studies from Aeronautical History*. Baltimore: Johns Hopkins Univ. Press, 1991.

[VT86] Alessandro Verri and V. Torre. Absolute depth estimate in stereopsis. *Journal of the Optical Society of America*, 3:297–299, 1986.

[VWS02] Roel Vertegaal, Ivo Weevers, and Changuk Sohn. GAZE-2: an attentive video conferencing system. In *CHI '02 extended abstracts on Human factors in Computing Systems*, CHI EA '02, pages 736–737, New York, NY, USA, 2002. ACM.

[WAH92] J. Weng, Narendra Ahuja, and Thomas S. Huang. Matching two perspective views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):806–825, 1992.

[Web08] WebM. VP8 Codec SDK. http://www.webmproject.org/tools/vp8-sdk/, 2008. Last accessed August 06, 2012.

[WG02] Peter Willemsen and Amy A. Gooch. Perceived egocentric distances in real, image-based, and traditional virtual environments. In *Proceedings of the IEEE Virtual Reality Conference 2002*, VR '02, pages 275–, Washington, DC, USA, 2002. IEEE Computer Society.

[Whe38] Charles Wheatstone. Contributions to the physiology of vision. part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, 128:371:394, 1838.

[WHM12] Julie Wagner, Stéphane Huot, and W.E. Mackay. BiTouch and BiPad: Designing Bimanual Interaction for Hand-held Tablets. In *CHI'12 - 30th International Conference on Human Factors in Computing Systems - 2012*, Austin, États-Unis, May 2012. ACM SIGCHI, ACM Press.

[Wil77] Ederyn Williams. Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin*, 84(5):963–976, 1977.

[Wil99] Willow Garage. OpenCV. http://opencv.willowgarage.com, 1999. Accessed July 15, 2012.

[WJK+13] Thomas Whelan, Hordur Johannsson, Michael Kaess, John J Leonard, and John McDonald. Robust real-time visual odometry for dense rgb-d mapping. In *IEEE International Conference on Robotics and Automation, ICRA*, 2013.

[WJV+05] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. *ACM Transaction on Graphics*, 24(3):765–776, July 2005.

[WKF+12] Thomas Whelan, Michael Kaess, Maurice Fallon, Hordur Johannsson, John Leonard, and John McDonald. Kintinuous: Spatially extended KinectFusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia, July 2012.

[WKLM13] Thomas Whelan, Michael Kaess, John J Leonard, and John McDonald. Deformation-based loop closure for large scale dense rgb-d slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2013.

[WLG04] Stephan Würmlin, Edouard Lamboray, and Markus H. Gross. 3D video fragments: dynamic point samples for real-time free-viewpoint video. *Computers Graphics*, 28:3–14, 2004.

[WMLS10] Daniel Wagner, Alessandro Mulloni, Tobias Langlotz, and Dieter Schmalstieg. Real-time panoramic mapping and tracking on mobile phones. In *Proceedings of the IEEE VR Conference*, pages 211–218. IEEE Computer Society, 2010.

[WRM+08] Robin Wolff, Dave Roberts, Alessio Murgia, Norman Murray, John Rae, Will Steptoe, Anthony Steed, and Paul Sharkey. Communicating eye gaze across a distance without rooting participants to the spot. In *Proceedings of the 2008 12th IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications*, DS-RT '08, pages 111–118, Washington, DC, USA, 2008. IEEE Computer Society.

[WSEK12] Christian Weissig, Oliver Schreer, Peter Eisert, and Peter Kauff. The ultimate immersive experience: Panoramic 3d video acquisition. In Klaus Schoeffmann, Bernard Merialdo, AlexanderG. Hauptmann, Chong-Wah Ngo, Yiannis Andreopoulos, and Christian Breiteneder, editors, *Advances in Multimedia Modeling*, volume 7131 of *Lecture Notes in Computer Science*, pages 671–681. Springer Berlin Heidelberg, 2012.

[WT09] Fuqu Wu and Melanie Tory. Photoscope: visualizing spatiotemporal coverage of photos for construction management. In *CHI*, pages 1103–1112, 2009.

[Wu07] Changchang Wu. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). http://cs.unc.edu/~ccwu/siftgpu, 2007. Accessed July 15, 2012.

[WZ08] Zeng-Fu Wang and Zhi-Gang Zheng. A region based stereo matching algorithm using cooperative optimization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[Xip00] Xiph.Org Foundation. Ogg vorbis. http://www.vorbis.com/, 2000. Accessed November 15, 2013.

[YCNB96] Kimiya Yamaashi, Jeremy R. Cooperstock, Tracy Narine, and William Buxton. Beating the limitations of camera-monitor mediated telepresence with extra eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 50–57, New York, NY, USA, 1996. ACM.

[YDMH99] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 215–224, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[YHMY12] H. Yamazoe, H. Habe, I. Mitsugami, and Y. Yagi. Easy depth sensor calibration. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 465–468, Nov 2012.

[YYDN07] Qingxiong Yang, Ruigang Yang, James Davis, and David Nister. Spatial-depth super resolution for range images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:1–8, 2007.

[ZCS02] Li Zhang, Brian Curless, and Steven M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *The 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 24–36, June 2002.

[ZDFL95] Zhengyou Zhang, Rachid Deriche, Olivier D. Faugeras, and Quang-Tuan Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1/2):87–119, November 1995.

[ZDS09] W. Zia, K. Diepold, and M. Sarkis. Optimization of video coding for telepresence applications. In *WACV09*, pages 1–8, 2009.

[ZF92] Zhengyou Zhang and Olivier D. Faugeras. Estimation of displacements from two 3-D frames obtained from stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(12):1141–1156, 1992.

[ZH04] Song Zhang and Peisen Huang. High-resolution, real-time 3D shape acquisition. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, volume 3, pages 28–36, Washington, DC, USA, 2004. IEEE Computer Society.

[Zha95] Zhengyou Zhang. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1222–1227, 1995.

[Zha97] Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–170, 1997.

[Zim04] Paul Michael Zimmons. *The influence of lighting quality on presence and task performance in virtual environments*. PhD thesis, The University of North Carolina at Chapel Hill, 2004.

[ZLF96] Zhengyou Zhang, Quang-Tuan Luong, and Olivier D. Faugeras. Motion of an uncalibrated stereo rig: Self-calibration and metric reconstruction. *IEEE Transactions on Robotics and Automation*, 12(1):103–113, 1996.

[ZLY10] Hong Zhang, Bo Li, and Dan Yang. Keyframe detection for appearance-based visual slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2010*, pages 2071–2076, October 2010.

[ZP03] Paul Zimmons and Abigail Panter. The influence of rendering quality on presence and task performance in a virtual environment. In *Virtual Reality, 2003. Proceedings. IEEE*, pages 293–294, 2003.

[ZSMG07] Zeev Zalevsky, Alexander Shput, Aviad Maizels, and Javier Garcia. Method and system for object reconstruction, 2007.

[ZTCS99] Ruo Zhang, Ping-Sing Tsai, J.E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690 –706, August 1999.

[ZVDWO10] Song Zhang, Daniel Van Der Weide, and James Oliver. Superfast phase-shifting method for 3-d shape measurement. *Optics Express*, 18(9):9684–9689, 2010.